Jacob Andreas

`jda@mit.edu`
`web.mit.edu/jda/www`

Employment

Massachusetts Institute of Technology

Associate Professor (pre-tenure), 2023–present.
X Consortium Career Development Assistant Professor, 2020–2023.
Assistant Professor, 2019–2020.

Microsoft

Principal Researcher, 2021–present.
Senior Researcher, 2018–2021.

Semantic Machines

Research Scientist, 2014–2018.

Education

University of California, Berkeley

Doctor of Philosophy, 2013–2018. Advisor: Dan Klein.

University of Cambridge

Master of Philosophy (with distinction), 2012–2013. Advisor: Stephen Clark.

Columbia University in the City of New York

Bachelor of Science (*summa cum laude*), 2008–2012. Thesis advisor: Michael Collins.

Fellowships, Awards & Honors

Alfred P. Sloan Research Fellowship, 2024
CAREER Award, National Science Foundation 2023
Kavli Fellow, National Academy of Sciences, 2022

Amazon Research Award, 2022
Sony Faculty Innovation Award, 2021
Samsung AI Researcher of the Year, 2021

Junior Bose Award for Teaching, MIT School of Engineering, 2023
Kolokotrones Education Award, MIT EECS, 2021

Best paper: *Visual grounding helps learn word meanings in low-data regimes*, NAACL 2024.
Area chair's award: *Compositionality as lexical symmetry*, ACL 2023.
Notable paper: *What learning algorithm is in-context learning?*, ICLR 2023.
Best paper honorable mention: *Modular multitask reinforcement learning with policy sketches*, ICML 2017.
Best paper: *Learning to compose neural networks for question answering*, NAACL 2016.

Facebook Graduate Fellowship, 2016–2018
National Science Foundation Graduate Fellowship, 2013–2016
M.Phil. Dissertation Prize, Computer Laboratory, Cambridge, 2013
Winston Churchill Scholarship, 2012–2013
Theodore R. Bashkow Prize (for computer science research), Columbia, 2012
Russell C. Mills Prize (for computer science coursework), Columbia, 2012
Tau Beta Pi, 2011
C. Prescott Davis Scholar, Columbia, 2008–2012

Preprints

(Available upon request.)

[1] *Bayesian preference elicitation with language models.*
Kunal Handa, Yarin Gal, Ellie Pavlick, Noah Goodman, Jacob Andreas, Alex Tamkin, Belinda Z. Li.

[2] *Eliciting human preferences with language models.*
Belinda Li*, Alex Tamkin*, Noah Goodman and Jacob Andreas.

[3] *Inspecting and editing knowledge representations in language models*.
Evan Hernandez, Belinda Li and Jacob Andreas.

[4] *Algorithmic capabilities of random transformers*.
Ziqian Zhong and Jacob Andreas.

[5] *Unforgettable generalization in language models*.
Eric Zhang, Leshem Choshen and Jacob Andreas.

[6] *An incomplete loop: Deductive, inductive, and abductive learning in large language models*.
Emmy Liu, Graham Neubig and Jacob Andreas.

[7] *Policy learning with a language bottleneck*.
Megha Srivastava, Cédric Colas, Dorsa Sadigh and Jacob Andreas.

[8] *Language modeling with editable external knowledge*.
Belinda Z. Li, Emmy Liu, Alexis Ross, Abbas Zeitoun, Graham Neubig and Jacob Andreas.

[9] *Language models trained on media diets can predict public opinion*.
Eric Chu, Jacob Andreas, Stephen Ansolabehere and Deb Roy.

[10] *Language-to-code translation with a single labeled example*.
Kaj Bostrom, Harsh Jhamtani, Hao Fang, Patrick Xia, Sam Thomson, Richard Shin, Benjamin Van Durme, Jason Eisner and Jacob Andreas.

Conference & Journal Papers

[11] *Toward in-context teaching: Adapting examples to students' misconceptions*.
Alexis Ross and Jacob Andreas.
ACL, 2024.

[12] *Deductive closure training of language models for coherence, accuracy and updatability*.
Afra Feyza Akyürek, Ekin Akyürek, Leshem Choshen, Derry Wijaya and Jacob Andreas.
ACL Findings, 2024.

[13] *Lexicon-level contrastive visual grounding improves language modeling*.
Chengxu Zhuang, Evelina Fedorenko and Jacob Andreas.
ACL Findings, 2024.

[14] *A multimodal automated interpretability agent*.
Tamar Rott Shaham*, Sarah Schwettmann*, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas and Antonio Torralba.
ICML, 2024.

[15] *In-context language learning: Architectures and algorithms*.
Ekin Akyürek, Bailin Wang, Yoon Kim and Jacob Andreas.
ICML, 2024.

[16] *Decomposing uncertainty for large language models through input clarification ensembling*.
Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang and Yang Zhang.
ICML, 2024.

[17] *Learning phonotactics from linguistic informants*.
Canaan Breiss*, Alexis Ross*, Amani Maina-Kilaas, Roger P. Levy and Jacob Andreas.
SCiL, 2024.

[18] *Natural language decomposition and interpretation of complex utterances*.
Harsh Jhamtani, Hao Fang, Patrick Xia, Eran Levy, Jacob Andreas and Benjamin Van Durme.
IJCAI, 2024.

[19] *Contextual and combinatorial structure in sperm whale vocalizations*.
Pratyusha Sharma, Shane Gero, Roger Payne, David F. Gruber, Daniela Rus*, Antonio Torralba*and Jacob Andreas*.
Nature Communications, 2024.

[20] *Visual grounding helps learn word meanings in low-data regimes*.
Chengxu Zhuang, Evelina Fedorenko and Jacob Andreas.
NAACL, 2024. (**Best paper**.)

[21] *Regularized conventions: Equilibrium computation as a model of pragmatic reasoning*.
Athul Paul Jacob, Gabriele Farina and Jacob Andreas.
NAACL, 2024.

[22] *Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual evaluations*.
Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas and Yoon Kim.
NAACL, 2024.

[23] *Interpreting user requests in the context of natural language standing instructions*.
Nikita Moghe, Patrick Xia, Jacob Andreas, Jason Eisner, Benjamin Van Durme and Harsh Jhamtani.
NAACL, 2024.

[24] *The consensus game: Language model generation via equilibrium search*.
Athul Paul Jacob, Yikang Shen, Gabriele Farina and Jacob Andreas.
ICLR, 2024. (**Spotlight presentation**.)

[25] *Linearity of relation decoding in transformer language models*.
Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov and David Bau.
ICLR, 2024. (**Spotlight presentation**.)

[26] *Learning adaptive planning representations with natural langauge guidance*.
Lionel Wong, Jiayuan Mao, Pratyusha Sharma, Zachary S Siegel, Jiahai Feng, Noa Korneev, Joshua B. Tenenbaum and Jacob Andreas.
ICLR, 2024.

[27] LILO*: Learning interpretable libraries by compressing and documenting code*.
Gabriel Grand, Lionel Wong, Matthew Bowers, Theo X. Olausson, Muxin Liu, Joshua B. Tenenbaum and Jacob Andreas.
ICLR, 2024.

[28] *Learning with language-guided state abstractions*.
Andi Peng, Ilia Sucholutsky, Belinda Z. Li, Theodore Sumers, Thomas L. Griffiths, Jacob Andreas and Julie Shah.
ICLR, 2024.

[29] *Modeling boundedly rational agents with latent inference budgets*.
Athul Paul Jacob, Abhishek Gupta, Jacob Andreas.
ICLR, 2024.

[30] *Decision-oriented dialogue for human–AI collaboration*.
Jessy Lin*, Nicholas Tomlin*, Jacob Andreas and Jason Eisner.
TACL, 2024.

[31] *Cognitive dissonance: Why do language model predictions disagree with their internal representations of truthfulness?*
Kevin Liu, Stephen Casper, Dylan Hadfield-Menell and Jacob Andreas.
EMNLP, 2023.

[32] *Encoding recursive computations in transformers with λ-layers*.
Shikhar Murthy, Pratyusha Sharma, Jacob Andreas and Christopher Manning.
EMNLP, 2023.

[33] *Alignment via mutual information*.
Shinjini Ghosh, Yoon Kim, Tahira Naseem, Ramon Fernandez Astudillo and Jacob Andreas.
CoNLL, 2023.

[34] *The clock and the pizza: Two stories in mechanistic explanation of neural networks*.
Ziming Liu*, Ziqian Zhong*, Max Tegmark and Jacob Andreas.
NeurIPS, 2023. (**Oral presentation.**)

[35] *A function interpretation benchmark for evaluating interpretability methods*.
Sarah Schwettmann, Tamar Rott Shaham, Joanna Materzynska, Neil Chowdhury, Shuang Li, Jacob Andreas, David Bau, Antonio Torralba.
NeurIPS Datasets & Benchmarks, 2023.

[36] *Lexical semantic content, not syntactic structure, is the main contributor to ANN-brain similarity of fMRI responses in the language network*.
Carina Kauf, Greta Tuckute, Roger Levy, Jacob Andreas and Evelina Fedorenko.
Neurobiology of Language, 2023.

[37] *Compositionality as lexical symmetry*.
Ekin Akyürek and Jacob Andreas.
ACL, 2023. (**Area chair's award**.)

[38] *Grokking of hierarchical structure in transformers*.
Shikhar Murty, Pratyusha Sharma, Jacob Andreas and Christopher Manning.
ACL, 2023.

[39] *Language modeling with latent situations*.
Belinda Z. Li, Max Nye and Jacob Andreas.
ACL Findings, 2023.

[40] *The whole truth and nothing but the truth: Faithful and controllable dialogue response generation with dataflow transduction and constrained decoding*.
Hao Fang*, Anusha Balakrishnan*, Harsh Jhamtani*, John Bufe, Jean Crawford, Jayant Krishnamurthy Adam Pauls, Jason Eisner, Jacob Andreas, Dan Klein.
ACL Findings, 2023.

[41] *Guiding pretraining in reinforcement learning with large language models*.
Yuqing Du*, Olivia Watkins*, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta and Jacob Andreas.
ICML, 2023.

[42] *PromptBoosting: Text classification with langauge models in ten forward passes*.
Bairu Hou, Joe O'Connor, Jacob Andreas, Yang Zhang and Shiyu Chang.
ICML, 2023.

[43] *What learning algorithm is in-context learning? Investigations with linear models*.
Ekin Akyürek, Dale Schuurmans, Jacob Andreas*, Tengyu Ma*and Denny Zhou*.
ICLR, 2023. (**Notable paper, oral presentation.**)

[44] *Characterizing intrinsic compositionality in transformers with tree projections*.
Shikhar Murty, Pratyusha Sharma, Jacob Andreas and Christopher Manning.
ICLR, 2023.

[45] *Language models as agent models*.
Jacob Andreas.
EMNLP Findings, 2022.

[46] *Tracing knowledge in language models back to the training data*.
Ekin Akyürek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas and Kelvin Guu.
EMNLP Findings, 2022.

[47] *Hierarchical phrase-based sequence-to-sequence learning*.
Bailin Wang, Ivan Titov, Jacob Andreas and Yoon Kim.
EMNLP, 2022.

[48] *Pre-trained language models for interactive decision-making*.
Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, Jacob Andreas, Igor Mordatch, Antonio Torralba, Yuke Zhu.
NeurIPS, 2022. (**Oral presentation.**)

[49] *Modeling strong and human-like gameplay with KL-regularized search*.
Athul Paul Jacob*, David Wu*, Gabriele Farina*, Adam Lerer, Hengyuan Hu, Anton Bakhtin, Jacob Andreas and Noam Brown.
ICML, 2022.

[50] *Toward understanding the communication in sperm whales*.
Jacob Andreas, Gašper Beguš, Michael M Bronstein, Roee Diamant, Denley Delaney, Shane Gero, Shafi Goldwasser, David F Gruber, Sarah de Haas, Peter Malkin, Roger Payne, Giovanni Petri, Daniela Rus, Pratyusha Sharma, Dan Tchernov, Pernille Tønnesen, Antonio Torralba, Daniel Vogt and Robert J Wood.
iScience, 2022.

[51] *Identifying concept libraries from language about object structure*.
Lio Wong*, William P. McCarthy*, Gabriel Grand*, Yoni Friedman, Joshua B. Tenenbaum, Jacob Andreas, Robert D. Hawkins, Judith E. Fan.
CogSci, 2022.

[52] *Correcting robot plans with natural language feedback*.
Pratyusha Sharma, Balakumar Sundaralingam, Valts Blukis, Chris Paxton, Tucker Hermans, Antonio Torralba, Jacob Andreas, Dieter Fox.
RSS, 2022.

[53] *Quantifying adaptability in pre-trained language models with 500 tasks*.
Belinda Z. Li, Jane Yu, Madian Khabsa, Luke Zettlemoyer, Alon Halevy and Jacob Andreas.
NAACL, 2022.

[54] *Skill induction and planning with latent language*.
Pratyusha Sharma, Antonio Torralba and Jacob Andreas.
ACL, 2022.

[55] *Natural language descriptions of deep visual features*.
Evan Hernandez, Sarah Schwettman, David Bau, Antonio Torralba and Jacob Andreas.
ICLR, 2022. (**Oral presentation.**)

[56] *Subspace regularizers for few-shot incremental classification*.
Afra Feyza Akyürek, Ekin Akyürek, Derry Wijaya and Jacob Andreas.
ICLR, 2022.

[57] *Teachable reinforcement learning via advice distillation*.
Olivia Watkins, Trevor Darrell, Pieter Abbeel, Jacob Andreas, Abhishek Gupta.
NeurIPS, 2021.

[58] *The low-dimensional linear geometry of contextualized word representations*.
Evan Hernandez and Jacob Andreas.
CoNLL, 2021.

[59] *How do neural sequence models generalize? Local and global context cues for out-of-distribution prediction*.
Anthony Bau and Jacob Andreas.
EMNLP, 2021.

[60] *Toward a visual concept vocabulary for generative adversarial networks*.
Sarah Schwettmann, Evan Hernandez, David Bau, Samuel Klein, Jacob Andreas, Antonio Torralba.
ICCV, 2021.

[61] *Leveraging natural language to learn program abstractions and search heuristics*.
Lio Wong, Kevin Ellis, Joshua B. Tenenbaum and Jacob Andreas.
ICML, 2021.

[62] *Implicit representations of meaning in neural language models.*
Belinda Z. Li, Max Nye and Jacob Andreas.
ACL, 2021.

[63] *Lexicon learning for few-shot sequence modeling.*
Ekin Akyürek and Jacob Andreas.
ACL, 2021.

[64] *What context features can transformer language models use?*
Joe O'Connor and Jacob Andreas.
ACL, 2021.

[65] *Value-agnostic conversational semantic parsing.*
Emmanouil Antonios Platanios, Adam Pauls, Subhro Roy, Yuchen Zhang, Alexander Kyte, Alan Guo, Sam Thomson, Jayant Krishnamurthy, Jason Wolfe, Jacob Andreas and Dan Klein.
ACL, 2021.

[66] *Multitasking inhibits semantic drift.*
Athul Paul Jacob, Mike Lewis and Jacob Andreas.
NAACL, 2021.

[67] *Compositional generalization for neural semantic parsing via span-level supervised attention.*
Pengcheng Yin, Hao Fang, Graham Neubig, Adam Pauls, Anthony Platanios, Yu Su, Sam Thomson and Jacob Andreas.
NAACL, 2021.

[68] *Representing partial programs with blended abstract semantics.*
Maxwell Nye, Yewen Pu, Matthew Bowers, Jacob Andreas, Joshua B. Tenenbaum, Armando Solar-Lezama.
ICLR, 2021.

[69] *Learning to recombine and resample data for compositional generalization.*
Ekin Akyürek, Afra Feyza Akyürek, and Jacob Andreas.
ICLR, 2021.

[70] *Compositional explanations of neurons.*
Jesse Mu and Jacob Andreas.
NeurIPS, 2020. (**Oral presentation.**)

[71] *A benchmark for systematic generalization in grounded language understanding.*
Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt and Brenden Lake.
NeurIPS, 2020.

[72] *Experience grounds language.*
Yonatan Bisk*, Ari Holtzman*, Jesse Thomason*, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto and Joseph Turian.
EMNLP, 2020.

[73] *Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment.*
Geeticka Chauhan, Ruizhi Liao, William Wells, Jacob Andreas, Xin Wang, Seth Berkowitz, Steven Horng, Peter Szolovits and Polina Golland.
MICCAI, 2020.

[74] *Task-oriented dialogue as dataflow synthesis.*
Semantic Machines et al.
TACL, 2020.

[75] *Good-enough compositional data augmentation*.
Jacob Andreas.
ACL, 2020.

[76] *A survey of reinforcement learning informed by natural language*.
Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, Tim Rocktäschel.
IJCAI, 2019.

[77] *Pragmatically informative text generation*.
Sheng Shen, Daniel Fried, Jacob Andreas and Dan Klein.
NAACL, 2019.

[78] *Measuring compositionality in representation learning*.
Jacob Andreas.
ICLR, 2019.

[79] *Guiding policies with language via meta-learning*.
John D Co-Reyes, Abhishek Gupta, Suvansh Sanjeev, Nick Altieri, Jacob Andreas, John DeNero, Pieter Abbeel, Sergey Levine.
ICLR, 2019.

[80] *Speaker–follower models for vision-and-language navigation*.
Daniel Fried*, Ronghang Hu*, Volkan Cirik*, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Trevor Darrell and Dan Klein.
NeurIPS, 2018.

[81] *Explainable neural computation via stack neural module networks*.
Ronghang Hu, Jacob Andreas, Kate Saenko and Trevor Darrell.
ECCV, 2018.

[82] *Can deep reinforcement learning solve Erdős–Selfridge–Spencer games?*
Maithra Raghu, Alex Irpan, Jacob Andreas, Robert Kleinberg, Quoc Le and Jon Kleinberg.
ICML, 2018.

[83] *Learning with latent language*.
Jacob Andreas, Dan Klein and Sergey Levine.
NAACL, 2018.

[84] *Unified pragmatic models for generating and following instructions*.
Daniel Fried, Jacob Andreas and Dan Klein.
NAACL, 2018.

[85] *Learning to reason: End to end module networks for visual question answering*.
Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell and Kate Saenko.
ICCV, 2017. (**Spotlight presentation.**)

[86] *Analogs of linguistic structure in deep representations*.
Jacob Andreas and Dan Klein.
EMNLP, 2017.

[87] *Modular multitask reinforcement learning with policy sketches*.
Jacob Andreas, Dan Klein and Sergey Levine.
ICML, 2017. (**Best paper honorable mention.**)

[88] *Translating neuralese*.
Jacob Andreas, Anca Dragan and Dan Klein.
ACL, 2017.

[89] *A minimal span-based constituency parser*.
Mitchell Stern, Jacob Andreas and Dan Klein.
ACL, 2017.

[90] *Modeling relationships in referential expressions with compositional modular networks*.
Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell and Kate Saenko.
CVPR, 2017. (**Spotlight presentation.**)

[91] *Reasoning about pragmatics with neural listeners and speakers*.
Jacob Andreas and Dan Klein.
EMNLP, 2016.

[92] *Learning to compose neural networks for question answering*.
Jacob Andreas, Marcus Rohrbach, Trevor Darrell and Dan Klein.
NAACL, 2016. (**Best paper.**)

[93] *Neural module networks*.
Jacob Andreas, Marcus Rohrbach, Trevor Darrell and Dan Klein.
CVPR, 2016. (**Oral presentation.**)

[94] *On the accuracy of self-normalized log-linear models*.
Jacob Andreas*, Maxim Rabinovich*, Dan Klein and Michael I. Jordan.
NeurIPS, 2015.

[95] *Alignment-based compositional semantics for instruction following*.
Jacob Andreas and Dan Klein.
EMNLP, 2015.

[96] *When and why are log-linear models self-normalizing?*
Jacob Andreas and Dan Klein.
NAACL, 2015.

[97] *Unsupervised transcription of piano music*.
Taylor Berg-Kirkpatrick, Jacob Andreas and Dan Klein.
NeurIPS, 2014. (**Spotlight presentation.**)

[98] *Grounding language with points and paths in continuous spaces*.
Jacob Andreas and Dan Klein.
CoNLL, 2014.

[99] *How much do word embeddings encode about syntax?*
Jacob Andreas and Dan Klein.
ACL, 2014.

[100] *Semantic parsing as machine translation*.
Jacob Andreas, Andreas Vlachos and Stephen Clark.
ACL, 2013.

[101] *Parsing graphs with hyperedge replacement grammars*.
David Chiang, Jacob Andreas, Daniel Bauer, Karl Moritz Hermann, Bevan Jones and Kevin Knight.
ACL, 2013.

[102] *Semantics-based machine translation with hyperedge replacement grammars*.
Bevan Jones*, Jacob Andreas*, Daniel Bauer*, Karl Moritz Hermann*, and Kevin Knight.
COLING, 2012.

[103] *Annotating agreement and disagreement in threaded discussion*.
Jacob Andreas, Sara Rosenthal and Kathleen McKeown.
LREC, 2012.

[104] *Fuzzy syntactic reordering for phrase-based statistical machine translation*.
Jacob Andreas, Nizar Habash and Owen Rambow.
WMT, 2011.

[105] *Semi-automated annotation for prepositional phrase attachment*.
Sara Rosenthal, William Lipovsky, Kathleen McKeown, Kapil Thadani and Jacob Andreas.
LREC, 2010.

Workshop Papers

[106] *From word models to world models: Translating from natural language to the language of thought*.
Lionel Wong, Gabriel Grand, Alexander K. Lew, Noah Goodman, Vikash Mansinghka, Jacob Andreas and Joshua B. Tenenbaum.
Society for Philosophy and Psychology, 2022. (**SPP William James Prize runner-up**.)

[107] *LaMPP: Language models as probabilistic priors for perception and action*.
Belinda Z. Li, Will Chang, Pratyusha Sharma and Jacob Andreas.
ICLR Workshop on Generative AI for Decision-Making, 2024.

[108] *ObSynth: An interactive synthesis system for generating object models from natural language specifications*.
Alex Gu, Tamara Mitrovska*, Daniela Velez*, Jacob Andreas and Armando Solar-Lezama.
ICLR Workshop on Deep Learning for Code, 2023.

[109] *Unnatural language processing: bridging the gap between synthetic and natural language data*.
Alana Marzoev, Sam Madden, Franz Kaashoek, Mike Cafarella and Jacob Andreas.
NeurIPS Workshop on Emergent Communication, 2019.

[110] *Learning to plan without a planner*.
Jacob Andreas, Mitchell Stern and Dan Klein.
NeurIPS Workshop on Neural Abstract Machines and Program Induction, 2016.

[111] *A generative model of vector space semantics*.
Jacob Andreas and Zoubin Ghahramani.
ACL Workshop on Continuous Vector Space Models and their Compositionality, 2013.

[112] *Detecting influencers in written online conversations*.
Or Biran, Sara Rosenthal, Jacob Andreas, Kathleen McKeown and Owen Rambow.
NAACL Workshop on Language and Social Media, 2012.

[113] *Corpus creation for new genres: a crowdsourced approach to PP attachment*.
Mukund Jha, Jacob Andreas, Kapil Thadani, Sara Rosenthal and Kathleen McKeown.
NAACL Workshop on Creating Speech and Language Data with Mechanical Turk, 2010.

Patents

[114] *Ambiguity resolution with dialogue search history*.
David Leo Wright Hall, David Burkett, Jesse Rusak, Jayant Krishnamurthy, Jason Wolfe, Adam Pauls, Alan Guo, Jacob Andreas, Daniel Klein.
U.S. Patent 11195523, 2022.

[115] *Generating dialogue events for natural language system*.
Jacob Andreas, Dan Klein, David Leo Wright Hall, Larry Gillick, Pengyu Chen..
U.S. Patent 11133001, 2022.

[116] *Response generation for conversational computing interface*.
Jacob Andreas, Jayant Sivarama Krishnamurthy, Alan Xinyu Guo, Andrei Vorobev, John Philip Bufe III, Jesse Daniel Eskes Rusak, Yuchen Zhang.
U.S. Patent 11410643, 2022.

[117] *Error recovery for conversational systems*.
David Leo Wright Hall, David Burkett, Jesse Rusak, Alexander Kolmykov-Zotov, Jason Wolfe, Jacob Andreas, Adam Pauls, John Bufe III, Jayant Krishnamurthy, Dan Klein.
U.S. Patent 11106536B2, 2021.

[118] *Natural language content generator*.
Jacob Andreas, David Leo Wright Hall, Dan Klein, Adam Pauls.
U.S. Patent US10713288B2, 2020.

[119] *Automated assistant for user interaction via speech*.
Jacob Andreas, Taylor Berg-Kirkpatrick, Charles Chen, Jordan Cohen, Laurence Gillick, David Leo Wright Hall, Dan Klein, Michael Newman, Adam Pauls, Daniel Roth, Jesse Rusak, Andrew Volpe, Steven Wegmann.
U.S. Patent 10276160B2, 2019.

Advising

Postdoctoral researchers

Leshem Choshen, 2023–2024
Anna Ivanova, 2022–2023 (with Ev Fedorenko, → psychology faculty at Georgia Tech)
Chengxu Zhuang, 2022–present (with Ev Fedorenko)
Cédric Colas, 2022–present (with Josh Tenenbaum)

Ph.D. students

Alexis Ross, 2022–present (NSF fellow)
Gabe Grand, 2021–present (with Josh Tenenbaum, NSF fellow)
Pratyusha Sharma, 2021–present (with Antonio Torralba)
Belinda Li, 2020–present (Claire Booth Luce and NDSEG fellow)
Ekin Akyürek, 2019–present (Amazon ScienceHub fellow)
Alana Marzoev, 2019–present (with Sam Madden and Franz Kaashoek; on leave as CEO of `ReadySet.io`)
Athul Paul Jacob, 2019–present
Lio Wong, Brain and Cognitive Sciences, MIT, 2019–present (with Josh Tenenbaum)
Evan Hernandez, 2019–2025

Ph.D. committees

Aviv Netanyahu, EECS, MIT, 2025
Will Brannon, Media Lab, MIT, 2025
Ben Lipkin, Brain and Cognitive Sciences, MIT, 2025
Felix Wang, EECS, MIT, 2024
Eric Lehman, EECS, MIT, 2024
Olivia Watkins, EECS, UC Berkeley, 2024
Ji Lin, EECS, MIT, 2023
Róbert Csórdas, IDSIA, 2023
Elias Stengel-Eskin, Johns Hopkins, 2023
Hao Zheng, Informatics, University of Edinburgh, 2023
Shuang Li, EECS, MIT, 2023
Sameer Khurana, EECS, MIT, 2023
Yilun Zhou, EECS, MIT, 2022
Tom McCoy, Cognitive Science, Johns Hopkins, 2022
Candace Ross, EECS, MIT, 2021
Max Nye, Brain and Cognitive Sciences, MIT, 2021
Eric Chu, Media Lab, MIT, 2021
Yu-An Chung, EECS, MIT, 2020
Dzmitry Bahdanau, Computer Science, University of Montreal, 2020

M.Eng. students

Eric Zhang, 2023–2025
Jason Lee, 2023–2024
Matthew Kearney, 2022–2023 (→ Ph.D. at Oxford, Rhodes Scholar)
Shinjini Ghosh, 2022–2023
Kevin Liu, 2022–2023
Anthony Bau, 2021–2022 (→ Philosophy M.A. at Tufts)
Joe O'Connor, 2019–present (→ Ph.D. at UCLA)

S.B. students

Reihaneh Iranmanesh, 2023
Muxin Liu, 2023
William Chen, 2022 (→ Ph.D. at U.C. Berkeley)
Noa Korneev, 2022
Jiahai Feng, 2022 (→ Ph.D. at U.C. Berkeley)
Nafisa Rashid, 2022 (→ Linguistics Ph.D. at U.C. Berkeley)
Josue Torres Fonseca, 2022 (→ Ph.D. at University of Michigan)
Anjali Kantharuban, 2021 (→ Ph.D. at CMU, Gates Cambridge Scholar)
Teona Baghashvili, 2021 (→ Ph.D. at Boston University)
Nitya Parthasarathy, 2020

Teaching

Seminar: Multi-Agent Communication (6.S893). MIT, 2023.
Natural Language Processing (6.861). MIT, 2020–present.
Fundamentals of Programming (6.101). MIT, 2023.
Unsupervised Machine Learning (short course). MIT Executive Education, 2022.

Seminar: Doing Things With Words (6.884). MIT, 2022.
Seminar: Neuro-symbolic Models for NLP (6.884). MIT, 2020.
Introduction to Machine Learning (6.036). MIT, 2019.
Artificial Intelligence (cs188). Berkeley, 2016.
Emerging Scholars Program (COMS 3998). Columbia, 2011.

Invited Talks, Panels, & Guest Lectures

Long-term visitor, Simons Institute Year on Language Models, 2024–2025.
Workshop: Understanding High-Level Intelligence
Workshop: Transformers as a Computational Model
MBL Summer School on Brains, Minds and Machines, August 2024.
***Keynote*, Society for Computation in Linguistics**, July 2024.
M200 Association, Jun 2024.
University College London, May 2024.
University of Oxford, May 2024.
UK AI Safety Institute, May 2024.
Cohere For AI Fireside Chat, January 2024.
University of Washington NLP Seminar, January 2024.

Columbia NLP Seminar, November 2023.
Keynote, Workshop on the Future of Academic Research in the Era of Large Pre-trained Models, University of Maryland, November 2023.
Keynote, Liberty Mutual Data Science Forum, October 2023.
Cornell Tech Learning Machines Seminar, October 2023.
House Permanent Select Committee on Intelligence, September 2023.
Harvard Center for Mathematical Sciences and Applications Big Data Conference, August 2023.
Anthropic, July 2023.
Break Through Tech AI Summer School, July 2023.
Institute for Artificial Intelligence and Fundamental Interactions Workshop on Language Models, July 2023.
CogSci Workshop on Discovering Abstractions, July 2023.
University of Edinburgh ILCC Seminar, July 2023.
CVPR Workshop on Machine Visual Common Sense, June 2023.
Berkeley Center for Human-Compatible AI Workshop, June 2023.
Simons Institute Workshop on Decoding Communication in Non-Human Species, June 2023.
Open Data Science Conference East, May 2023.
Koç University AI Lab Distinguished Seminar Series, May 2023.
NYU Text as Data Seminar, April 2023.
Sony, April 2023.
AAAI Workshop on Representation Learning for Human-Compatible AI, February 2023.

Weizmann Institute Workshop on Language Models for Code, December 2022.
CoRL Workshop on Language and Robotics, December 2022.
CoRL Workshop on Aligning Human and Robot Representations, December 2022.
University of Chicago NLP Seminar, November 2022.
EMNLP Workshop on Learning-to-Learn Thorugh Interaction, November 2022.
University of Pennsylvania NLP seminar, October 2022.
Momentum AI Summer School, August 2022.
Panelist, CogSci Discussion on Neural Network Models of Cognition, July 2022.
***Keynote*, Conference on Lexical and Computational Semantics**, July 2022.

***Keynote*, Summer School on Neuro-Symbolic Learning**, July 2022.
Stanford NLP seminar, July 2022.
Invited Tutorial at RLDM, June 2022.
Johns Hopkins NLP seminar, June 2022.
ICLR Workshop on Deep Learning for Code, April 2022.
IBM Workshop on Unifying Statistical and Symbolic AI, January 2022.

NeurIPS Workshop on Explainable AI for Debugging and Diagnosis, December 2021.
Guest Lecture in Advanced Topics in Learning & Decisionmaking, UC Berkeley, November 2021.
Guest Lecture in Deep Learning, MIT, November 2021.
EMNLP Workshop on Sustainable NLP, November 2021.
Samsung AI Forum, November 2021.
Carnegie Mellon Language Technologies Institute Colloquium, September 2021.
University of Osnabrück Workshop on Computational Cognition, September 2021.
Princeton NLP Seminar, August 2021.
WING Lab Seminar, National University of Singapore, July 2021.
NAACL Workshop on Advances in Language and Vision Research, June 2021.
Liberty Mutual, June 2021.
MIT–IBM Watson AI Lab, June 2021.
GdR LIFT Seminar, University of Paris, June 2021.
Google Workshop on Conceptual Understanding of Deep Learning, May 2021.
UMass Lowell Computer Science Department Colloquium, April 2021.
NSF Expeditions Seminar: Understanding the World Through Code, March 2021.
DARK Lab seminar, UCL, March 2021.
Naval Research Laboratory, February 2021.
Samsung Workshop on Deep Learning and Logic, February 2021.

***Plenary Panelist*, COLING**, December 2020.
Self-Organizing Conference on Machine Learning, December 2020.
MIT Embodied Intelligence Seminar, November 2020.
EMNLP Workshop on Interactive and Executable Semantic Parsing, November 2020.
University of Edinburgh, November 2020.
Polytechnique Montreal / MILA Department Colloquium, October 2020.
Open Data Science Conference West, October 2020.
Simons Institute Workshop on Deep Reinforcement Learning, September 2020.
The Thesis Review podcast, September 2020.
Hazy Lab Seminar, Stanford, August 2020.
Apple Workshop on Natural Language Understanding, August 2020.
ACL Workshop on Conversational AI, July 2020.

Panelist, NeurIPS Workshop on Context & Compositionality in Artificial and Neural Systems, Dec. 2019.
NeurIPS Workshop on Emergent Communication, December 2019.
UT Austin AI Lectures, November 2019.
Open Data Science Conference West, October 2019.
ICML Workshop on Multi-task RL, June 2019.
Re·work Deep RL Summit, June 2019.
North Carolina State University ECE Interdisciplinary Distinguished Seminar Series, May 2019.
University of Tel Aviv Distinguished Lecture Series, May 2019.

NeurIPS Symposium on Deep Reinforcement Learning, December 2018.
NYU Text as Data Seminar, November 2018.

Microsoft Research, September 2018.
CLASP Workshop on Dialogue and Perception, June 2018.
DeepMind, June 2018.
CVPR Workshop on Visual Question Answering and Visual Dialogue, June 2018.
UC Berkeley, April 2018.
MIT, April 2018.
Stanford, April 2018.
Georgia Tech, March 2018.
University of Pennsylvania, March 2018.
Carnegie Mellon, March 2018.
Columbia, February 2018.
University of Montreal, February 2018.
McGill, February 2018.
TTI Chicago, January 2018.
Society for Computation in Linguistics, January 2018.

Panelist, NeurIPS Workshop on Emergent Communication, December 2017.
Allen Institute for AI, October 2017.
University of Washington, October 2017.
Microsoft Research, October 2017.
Facebook Fellows Workshop, September 2017.
AI2 *NLP Highlights* podcast, September 2017.
University of Amsterdam, September 2017.
Stanford, September 2017.
MIT, February 2017.
Harvard, February 2017.

Google Research, September 2016.
TTI Chicago, September 2016.

Berkeley Workshop on Algorithms for Human–Robot Interaction, November 2015.
Berkeley Center for New Music and Audio Technology, April 2015.

Selected media coverage

Improving truthfulness and coherence in LMs:

- Quanta: *Game Theory Can Make AI More Correct and Efficient* , May 2024.

Understanding deep networks:

- Quanta: *How Do Machines Grok Data?* April 2024.
- Scientific American: *How AI Knows Things No One Told It*, May 2023.
- Motherboard: *Scientists Made a Mind-Bending Discovery About How AI Actually Works*, February 2023.

AI for scientific discovery:

- The New York Times: *Scientists Find an 'Alphabet' in Whale Songs*, May 2024.
- New Yorker: *Can We Talk to Whales?* September 2023.
- The Atlantic: *AI Is Unlocking the Human Brain's Secrets*, May 2023.

Grants

National Science Foundation (CAREER). Learning structured models with natural language supervision. 2023–2027.

Liberty Mutual. Grounding Language models in domain knowledge. 2023–2024.

National Science Foundation (Medium). Bootstrapping natural feedback for reinforcement learning. With Abhishek Gupta. 2022–2025.

National Science Foundation (Large). Combining Learning and Formal Verification for Scalable Machine Programming. With Armando Solar-Lezama, Saman Amarasinghe, Jonathan Ragan-Kelly and Adam Chlipala. 2022-2025.

OpenPhilanthropy Foundation. Grounding language model behavior in latent representations of communicative intent. With Dylan Hadfield-Menell. 2022–2023.

Liberty Mutual. Interpretable AI for model outputs. 2022–2023.

Amazon Research Award. Natural language summaries of deep networks and decisions. 2022.

MIT Quest for Intelligence. Representations of meaning in minds and machines. With Ev Fedorenko, Roger Levy, and Athulya Aravind. 2021.

Sony Faculty Innovation Award. Natural language summaries of deep features and decisions. 2021–2022.

Systems That Learn @ CSAIL. Human- and machine-generated descriptions of deep features. 2021.

Machine Learning Applications @ CSAIL. Deep compositional models for multilingual language processing. 2021.

TED Audacious Prize. Project CETI: The Cetacean Translation Initiative. With Gašper Beguš, Michael M. Bronstein, Roee Diamant, Shane Gero, Shafi Goldwasser, David Gruber, Roger Payne, Giovanni Petri, Daniela Rus, Dan Tchernov, Antonio Torralba, and Robert Wood. 2020–2024.

IBM. Building bridges towards universal HLT: incorporating linguistic structure into neural models. With Regina Barzilay, Jim Glass, Roger Levy and Tommi Jaakkola. 2021–2024.

J.H. and E.V. Wade Fund. Decoding the language of whales. 2020–2021.

Takeda Pharmaceuticals. Automating medical review assessment with machine generated rationales. With Regina Barzilay and Tommi Jaakkola. 2020–2022.

NVIDIA. Understanding and manipulating images with natural language guidance. 2019–2020.

Professional Activities & Service

Campus

Artificial intelligence + Decision-Making Curriculum Committee, 2020–present.
Quest for Intelligence Advisory Committee, 2023–present.
Sprowls Dissertation Award Committee, 2020.
Dean’s action group on Social & Ethical Responsibilities of Computing, 2020.
President, Berkeley CS Graduate Student Association, 2014–2015.
Programming coach, 2Train Robotics (FIRST 395), 2010–2012.

Research Community

Panelist: National Science Foundation
Action Editor: ACL Rolling Review
Senior Area Chair: NeurIPS
Area Chair: ACL, EMNLP, CoNLL, NeurIPS, ICML
Standing Reviewer: TACL
Program Committee: ACL*, NAACL, EMNLP, EACL, NeurIPS*, ICML, ICLR*, UAI, CVPR, SCiL.
(*outstanding reviewer award)
Ad hoc reviewer: Review of Philosophy and Psychology, JMLR, PAMI, SIGCOMM
Organizing Committee:
ACL 2022 Workshop on Learning with Natural Language Supervision
ICML 2020 Workshop on Language and Reinforcement Learning
ICML 2019 Workshop on Adaptive and Multitask Learning
NAACL 2019 Workshop on Spatial Language Understanding and Language Grounding for Robotics
ACL 2017 Workshop on Language Grounding for Robotics
NAACL 2016 Student Research Workshop
Senior Advisory Committee:
NeurIPS 2022 Workshop on Language and Reinforcement Learning

Et cetera

UC Berkeley Chamber Chorus, 2014–2019.
Cambridge University Music Society Chorus, 2013.
Churchill College Boat Club, 2012–2013.
Lifetime member & full member, Philolexian Society, 2012.
Eagle scout, 2008.

Last updated June 21, 2024.