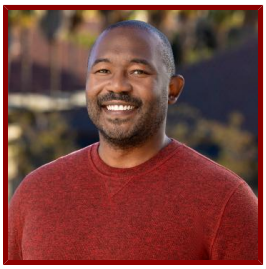
Stanford



Sanmi Koyejo

Assistant Professor of Computer Science

Bio

BIO

Sanmi Koyejo is an Assistant Professor in the Department of Computer Science at Stanford University and an adjunct Associate Professor at the University of Illinois at Urbana-Champaign. He leads the Stanford Trustworthy AI Research (STAIR) lab, which develops measurement-theoretic foundations for trustworthy AI systems, spanning AI evaluation science, algorithmic accountability, and privacy-preserving machine learning, with applications to healthcare and scientific discovery. His research on AI capabilities evaluation has challenged conventional understanding in the field, including work on measurement frameworks cited in the 2024 Economic Report of the President.

Koyejo has received the Presidential Early Career Award for Scientists and Engineers (PECASE), Skip Ellis Early Career Award, Alfred P. Sloan Research Fellowship, NSF CAREER Award, and multiple outstanding paper awards at flagship venues, including NeurIPS and ACL. He has delivered keynote presentations at major conferences, including ECCV and FAccT. He serves in key leadership roles, including Board President of Black in AI, Board of Directors of the Neural Information Processing Systems Foundation, and other leadership positions in professional organizations advancing AI research and broadening participation in the field.

ACADEMIC APPOINTMENTS

- Assistant Professor, Computer Science
- Member, Bio-X
- Member, Wu Tsai Human Performance Alliance
- Member, Wu Tsai Neurosciences Institute

Teaching

COURSES

2025-26

- AI Measurement Science: CS 321M (Spr)
- Governing Artificial Intelligence: Law, Policy, and Institutions: COMM 152A, COMM 252A, CS 283, GLOBAL 245B, INTLPOL 245B (Aut)
- Governing Artificial Intelligence: Law, Policy, and Institutions: LAW 4052 (Aut)
- Governing Artificial Intelligence: Law, Policy, and Institutions: POLISCI 145B, POLISCI 445B (Aut)
- Machine Learning: CS 229, STATS 229 (Win)
- Machine Learning from Human Preferences: CS 329H (Aut)

2024-25

- Machine Learning: CS 229, STATS 229 (Win)
- Machine Learning from Human Preferences: CS 329H (Aut)

2023-24

- Artificial Intelligence: Principles and Techniques: CS 221 (Spr)
- Machine Learning: CS 229, STATS 229 (Win)
- Machine Learning from Human Preferences: CS 329H (Aut)

2022-23

- Artificial Intelligence: Principles and Techniques: CS 221 (Spr)

STANFORD ADVISEES

Doctoral Dissertation Reader (AC)

Edward Chen, Richard Chen, Kanishk Gandhi

Orals Chair

Ravi Sojitra

Postdoctoral Faculty Sponsor

Youssef Allouah, Alexander Spangher, Zeyu Tang

Doctoral Dissertation Advisor (AC)

Steven Dillmann, Zach Robertson

Orals Evaluator

Kanishk Gandhi, Maya Varma

Doctoral Dissertation Co-Advisor (AC)

Ahmed Ahmed, Suhana Bedi, Fangrui Huang, Josh Kazdan, Alisa Levin, Kara Liu, Ken Liu, Anka Reuel, Neha Srivathsa, Alyssa Unell, Maya Varma

Master's Program Advisor

Stefan Ene, Nicolás Kennedy, Sreyana Kukadia, Hoang Nguyen, Isaac Park, Nestor Perez Fernandez, Jacob Rubenstein, Haoyue Xiao, Sean Yoon, Christine Zhang

Postdoctoral Research Mentor

Joachim Baumann

Doctoral (Program)

Nicole Chiou, Natalie Dullerud, Rui Li, Brando Miranda, Zach Robertson, Rylan Schaeffer, Nikil Selvam, Colin Sullivan, Sang Truong, Yibo Zhang

Publications

PUBLICATIONS

- **Holistic evaluation of large language models for medical tasks with MedHELM.** *Nature medicine*
  Bedi, S., Cui, H., Fuentes, M., Unell, A., Wornow, M., Banda, J. M., Kotecha, N., Keyes, T., Mai, Y., Oez, M., Qiu, H., Jain, S., Schettini, et al
  2026
- **Shaping AI's Impact on Billions of Lives** *COMMUNICATIONS OF THE ACM*
  Cuellar, M., Dean, J., Doshi-Velez, F., Hennessy, J., Konwinski, A., Koyejo, S., Moiloa, P., Pierson, E., Patterson, D.
  2026; 69 (1): 54-65

- **The inadequacy of offline large language model evaluations: A need to account for personalization in model behavior.** *Patterns (New York, N.Y.)*
  Wang, A., Ho, D. E., Koyejo, S.
  2025; 6 (12): 101397

- **TIMER: temporal instruction modeling and evaluation for longitudinal clinical records.** *NPJ digital medicine*
  Cui, H., Unell, A., Chen, B., Fries, J. A., Alsentzer, E., Koyejo, S., Shah, N. H.
  2025; 8 (1): 577

- **Advancing science- and evidence-based AI policy.** *Science (New York, N.Y.)*
  Bommasani, R., Arora, S., Chayes, J., Choi, Y., Cuéllar, M. F., Fei-Fei, L., Ho, D. E., Jurafsky, D., Koyejo, S., Lakkaraju, H., Narayanan, A., Nelson, A., Pierson, et al
  2025; 389 (6759): 459-461

- **Rethinking machine unlearning for large language models** *NATURE MACHINE INTELLIGENCE*
  Liu, S., Yao, Y., Jia, J., Casper, S., Baracaldo, N., Hase, P., Yao, Y., Liu, C., Xu, X., Li, H., Varshney, K. R., Bansal, M., Koyejo, et al
  2025

- **Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review.** *JAMA*
  Bedi, S., Liu, Y., Orr-Ewing, L., Dash, D., Koyejo, S., Callahan, A., Fries, J. A., Wornow, M., Swaminathan, A., Lehmann, L. S., Hong, H. J., Kashyap, M., Chaurasia, et al
  2024

- **Crossing Linguistic Horizons: Finetuning and Comprehensive Evaluation of Vietnamese Large Language Models**
  Truong, S. T., Nguyen, D. Q., Toan Nguyen, Le, D. D., Truong, N. N., Tho Quan, Koyejo, S.
  edited by Duh, K., Gomez, H., Bethard, S.
  ASSOC COMPUTATIONAL LINGUISTICS-ACL.2024: 2849-2900

- **The inadequacy of offline large language model evaluations: A need to account for personalization in model behavior** *PATTERNS*
  Wang, A., Ho, D. E., Koyejo, S.
  2025; 6 (12)

- **Evaluating anti-LGBTQIA+ medical bias in large language models.** *PLOS digital health*
  Chang, C. T., Srivathsa, N., Bou-Khalil, C., Swaminathan, A., Lunn, M. R., Mishra, K., Koyejo, S., Daneshjou, R.
  2025; 4 (9): e0001001

- **Fidelity of Medical Reasoning in Large Language Models.** *JAMA network open*
  Bedi, S., Jiang, Y., Chung, P., Koyejo, S., Shah, N.
  2025; 8 (8): e2526021

- **Advancing oil and gas emissions assessment through large language model data extraction** *ENERGY AND AI*
  Chen, Z., Zhong, R., Long, W., Tanga, H., Wang, A., Liu, Z., Yang, X., Ren, B., Littlefield, J., Koyejo, S., Masnadi, M. S., Brandt, A. R.
  2025; 20

- **The Reality of AI and Biorisk**
  Peppin, A., Reuel, A., Casper, S., Jones, E., Strait, A., Anwar, U., Agrawal, A., Kapoor, S., Koyejo, S., Pellat, M., Bommasani, R., Frosst, N., Hooker, et al
  ASSOC COMPUTING MACHINERY.2025: 763-771

- **Logits are All We Need to Adapt Closed Models**
  Hiranandani, G., Wu, H., Mukherjee, S., Koyejo, S.
  edited by Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., Zhu, J.
  JMLR-JOURNAL MACHINE LEARNING RESEARCH.2025: 23261-23289

- **Collapse or Thrive? Perils and Promises of Synthetic Data in a Self-Generating World**
  Kazdan, J., Schaeffer, R., Dey, A., Gerstgrasser, M., Rafailov, R., Donoho, D., Koyejo, S.
  edited by Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., Zhu, J.
  JMLR-JOURNAL MACHINE LEARNING RESEARCH.2025: 29469-29494

- **Decision from Suboptimal Classifiers: Excess Risk Pre- and Post-Calibration**

Perez-Lebel, A., Varoquaux, G., Koyejo, S., Doutreligne, M., Le Morvan, M.
edited by Li, Y., Mandt, S., Agrawal, S., Khan, E.
JMLR-JOURNAL MACHINE LEARNING RESEARCH.2025

- **Fairness through Difference Awareness: Measuring Desired Group Discrimination in LLMs**
  Wang, A., Phan, M., Ho, D. E., Koyejo, S.
  edited by Che, W., Nabende, J., Shutova, E., Pilehvar, M. T.
  ASSOC COMPUTATIONAL LINGUISTICS-ACL.2025: 6867-6893

- **Riding on the Back of a Whale: A Hackathon Framework for Introducing High School Students to Large Language Models**
  Nguyen, D., Le, D., Nguyen, L., Vu, Q., Le, T., Nguyen, D., Huynh, N., Nguyen, H., Tran, P., Le, D., Truong, S., Koyejo, S., Lea, et al
  edited by Cristea, A. I., Walker, E., Lu, Y., Santos, O. C., Isotani, S.
  SPRINGER INTERNATIONAL PUBLISHING AG.2025: 201-209

- **More than Marketing? On the Information Value of AI Benchmarks for Practitioners**
  Hardy, A., Reuel, A., Meimandi, K., Soder, L., Griffith, A., Asmar, D. M., Koyejo, S., Bernstein, M. S., Kochenderfer, M., ACM
  ASSOC COMPUTING MACHINERY.2025: 1032-1047

- **Publisher Correction: Increasing the presence of BIPOC researchers in computational science.** *Nature computational science*
  Chen, C. Y., Christoffels, A., Dube, R., Enos, K., Gilbert, J. E., Koyejo, S., Leigh, J., Liquido, C., McKee, A., Noe, K., Peng, T., Taiuru, K.
  2024

- **Increasing the presence of BIPOC researchers in computational science.** *Nature computational science*
  Chen, C. Y., Christoffels, A., Dube, R., Enos, K., Gilbert, J. E., Koyeji, S., Leigh, J., Liquido, C., McKee, A., Noe, K., Peng, T. Q., Taiuru, K.
  2024; 4 (9): 646-653

- **Artificial Intelligence, Social Responsibility, and the Roles of the University** *COMMUNICATIONS OF THE ACM*
  Bosch, N., Chan, A., Davis, J. L., Gutierrez, R., He, J., Karahalios, K., Koyejo, S., Loui, M. C., Mendenhall, R., Sanfilippo, M., Tong, H., Varshney, L. R., Wang, et al
  2024; 67 (8): 22-25

- **Single-Trial Detection and Classification of Event-Related Optical Signals for a Brain-Computer Interface Application.** *Bioengineering (Basel, Switzerland)*
  Chiou, N., Günal, M., Koyejo, S., Perpetuini, D., Chiarelli, A. M., Low, K. A., Fabiani, M., Gratton, G.
  2024; 11 (8)

- **Bridging gaps in automated acute myocardial infarction detection between high-income and low-income countries.** *PLOS global public health*
  Chiou, N., Koyejo, S., Ngaruiya, C.
  2024; 4 (6): e0003240

- **Bridging gaps in automated acute myocardial infarction detection between high-income and low-income countries** *PLOS GLOBAL PUBLIC HEALTH*
  Chiou, N., Koyejo, S., Ngaruiya, C.
  2024; 4 (6): e0003240

- **Author Correction: Opportunistic detection of type 2 diabetes using deep learning from frontal chest radiographs.** *Nature communications*
  Pyrros, A., Borstelmann, S. M., Mantravadi, R., Zaiman, Z., Thomas, K., Price, B., Greenstein, E., Siddiqui, N., Willis, M., Shulhan, I., Hines-Shah, J., Horowitz, J. M., Nikolaidis, et al
  2024; 15 (1): 4817

- **Impact of biased models in the context of fairness towards patients, and how to avoid or minimise biases in our datasets**
  Koyejo, S.
  ELSEVIER IRELAND LTD.2024: S46

- **Latent Multimodal Functional Graphical Model Estimation** *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION*
  Tsai, K., Zhao, B., Koyejo, S., Kolar, M.
  2024; 119 (547): 2217-2229

- **The Case for Globalizing Fairness: A Mixed Methods Study on Colonialism, AI, and Health in Africa**
  Asiedu, M., Dieng, A., Haykel, I., Rostamzadeh, N., Pfohl, S., Nagpal, C., Nagawa, M., Oppong, A., Koyejo, S., Heller, K., ACM

ASSOC COMPUTING MACHINERY.2024

- **Adaptive Compression in Federated Learning via Side Information**
  Isik, B., Pase, F., Gunduz, D., Koyejo, S., Weissman, T., Zorzi, M.
  edited by Dasgupta, S., Mandt, S., Li, Y.
  JMLR-JOURNAL MACHINE LEARNING RESEARCH.2024

- **Invariant Aggregator for Defending against Federated Backdoor Attacks**
  Wang, X., Dimitriadis, D., Koyejo, S., Tople, S.
  edited by Dasgupta, S., Mandt, S., Li, Y.
  JMLR-JOURNAL MACHINE LEARNING RESEARCH.2024

- **Causally Inspired Regularization Enables Domain General Representations**
  Salaudeen, O., Koyejo, S.
  edited by Dasgupta, S., Mandt, S., Li, Y.
  JMLR-JOURNAL MACHINE LEARNING RESEARCH.2024

- **Proxy Methods for Domain Adaptation**
  Tsai, K., Pfohl, S. R., Salaudeen, O., Chiou, N., Kusner, M. J., DAmour, A., Koyejo, S., Gretton, A.
  edited by Dasgupta, S., Mandt, S., Li, Y.
  JMLR-JOURNAL MACHINE LEARNING RESEARCH.2024

- **Towards Trustworthy Large Language Models**
  Koyejo, S., Li, B., Assoc computing machinery
  ASSOC COMPUTING MACHINERY.2024: 1126-1127

- **Bayesian Optimization for Crop Genetics with Scalable Probabilistic Models**
  Azam, R., Truong, S. T., Fernandes, S. B., Leakey, A. D. B., Lipka, A., El-Kebir, M., Koyejo, S.
  edited by Antoran, J., Naesseth, C. A.
  JMLR-JOURNAL MACHINE LEARNING RESEARCH.2024: 30-44

- **Disentangling Fact from Grid Cell Fiction in Trained Deep Path Integrators.** *ArXiv*
  Schaeffer, R., Khona, M., Koyejo, S., Fiete, I. R.
  2023

- **Longitudinal assessment of demographic representativeness in the Medical Imaging and Data Resource Center open data commons** *JOURNAL OF MEDICAL IMAGING*
  Whitney, H. M., Baughan, N., Myers, K. J., Drukker, K., Gichoya, J., Bower, B., Chen, W., Gruszauskas, N., Kalpathy-Cramer, J., Koyejo, S., Sa, R. C., Sahiner, B., Zhang, et al
  2023; 10 (6): 61105

- **Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment** *JOURNAL OF MEDICAL IMAGING*
  Drukker, K., Chen, W., Gichoya, J., Gruszauskas, N., Kalpathy-Cramer, J., Koyejo, S., Myers, K., Sa, R. C., Sahiner, B., Whitney, H., Zhang, Z., Giger, M.
  2023; 10 (6): 061104

- **Opportunistic detection of type 2 diabetes using deep learning from frontal chest radiographs.** *Nature communications*
  Pyrros, A., Borstelmann, S. M., Mantravadi, R., Zaiman, Z., Thomas, K., Price, B., Greenstein, E., Siddiqui, N., Willis, M., Shulhan, I., Hines-Shah, J., Horowitz, J. M., Nikolaidis, et al
  2023; 14 (1): 4039

- **Fast Optical Signals for Real-Time Retinotopy and Brain Computer Interface.** *Bioengineering (Basel, Switzerland)*
  Perpetuini, D., Gunal, M., Chiou, N., Koyejo, S., Mathewson, K., Low, K. A., Fabiani, M., Gratton, G., Chiarelli, A. M.
  2023; 10 (5)

- **One Policy is Enough: Parallel Exploration with a Single Policy is Near-Optimal for Reward-Free Reinforcement Learning**
  Cisneros-Velarde, P., Lyu, B., Koyejo, S., Kolar, M.
  edited by Ruiz, F., Dy, J., VanDeMeent, J. W.
  JMLR-JOURNAL MACHINE LEARNING RESEARCH.2023

- **Finite-sample Guarantees for Nash Q-learning with Linear Function Approximation**
  Cisneros-Velarde, P., Koyejo, S.
  edited by Evans, R. J., Shpitser
  JMLR-JOURNAL MACHINE LEARNING RESEARCH.2023: 424-432

- **Unraveling the Connections between Privacy and Certified Robustness in Federated Learning Against Poisoning Attacks**
  Xie, C., Long, Y., Chen, P., Li, Q., Koyejo, S., Li, B., ACM
  ASSOC COMPUTING MACHINERY.2023: 1511-1525

- **Self-Supervised Learning of Representations for Space Generates Multi-Modular Grid Cells**
  Schaeffer, R., Khona, M., Ma, T., Eyzaguirre, C., Koyejo, S., Fiete, I.
  edited by Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S.
  NEURAL INFORMATION PROCESSING SYSTEMS (NIPS).2023

- **DECODINGTRUST: A Comprehensive Assessment of Trustworthiness in GPT Models**
  Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., Truong, S. T., Arora, S., Mazeika, et al
  edited by Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S.
  NEURAL INFORMATION PROCESSING SYSTEMS (NIPS).2023

- **Pairwise Ranking Losses of Click-Through Rates Prediction for Welfare Maximization in Ad Auctions**
  Lyu, B., Feng, Z., Robertson, Z., Koyejo, S.
  edited by Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J.
  JMLR-JOURNAL MACHINE LEARNING RESEARCH.2023

- **Adapting to Latent Subgroup Shifts via Concepts and Proxies**
  Alabdulmohsin, I., Chiou, N., D'Amour, A., Gretton, A., Koyejo, S., Kusner, M. J., Pfohl, S. R., Salaudeen, O., Schrouff, J., Tsai, K.
  edited by Ruiz, F., Dy, J., VanDeMeent, J. W.
  JMLR-JOURNAL MACHINE LEARNING RESEARCH.2023

- **Fair Wrapping for Black-box Predictions**
  Soen, A., Alabdulmohsin, I., Koyejo, S., Mansour, Y., Moorosi, N., Nock, R.
  edited by Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A.
  NEURAL INFORMATION PROCESSING SYSTEMS (NIPS).2022