Schwarzman Scholar drawing lessons from China's economic rise for developing countries. With a keen interest in public policy, her research interests are in building machine learning systems that work effectively in resource-constrained contexts for developing countries.

- Email: jquaye@g.harvard.edu
- Web page: https://www.linkedin.com/in/jessicaquaye/

Juan Ciro is a Software Developer at MLCommons, responsible for leading the development of the innovative Dynabench platform. He holds a degree in Engineering and a Master's degree in Artificial Intelligence, with a focus on Deep Learning, from the International University of Applied Science. With over six years of experience in software development and research, Juan has made significant contributions to the field of machine learning, including the creation of open source datasets such as Multilingual Spoken Words and People's Speech, which was presented at NeurIPS 2022, a renowned conference in the field.

- Email: juanciro@mlcommons.org
- Web page: https://www.linkedin.com/in/juan-manuel-ciro-torres-471015aa/

Lora Aroyo is a Research Scientist at Google, NYC, where she works on research for Data Excellence by specifically focusing on metrics to measure quality of human-labeled data in a reliable and transparent way. She was one of the core organizers of the first data-centric workshop at NeurIPS2021 and led the efforts for the adversarial CATS4ML challenge. Lora is a co-chair of the HCOMP steering committee for the AAAI Human Computation conference and a president of the User Modeling community UM Inc, which serves as a steering committee for the ACM Conference Series "User Modeling, Adaptation and Personalization" (UMAP) sponsored by SIGCHI and SIGWEB. She is also a member of the ACM SIGCHI conferences board. Prior to joining Google, Lora was a computer science professor at the VU University Amsterdam. Dr. Aroyo has been conference chair, PC chair or track chairs for more than 10 conferences and has organized more than 20 workshops and tutorials in the area of Data Quality and Reliability, Human Computation, User Modeling and Semantic Web.

- Email: l.m.aroyo@gmail.com
- Web page: http://lora-aroyo.org
- Google Scholar: https://scholar.google.com/citations?user=FXGgl5IAAAAJ

Max Bartolo is a researcher at Cohere and a final-year PhD student with the UCL NLP group under the supervision of Pontus Stenetorp and Sebastian Riedel. His research lies at the intersection of model robustness and dynamic adversarial data collection, and he is a co-creator of Dynabench. Max co-organized the Dynamic Adversarial Data Collection (DADC) workshop at NAACL 2022 and the Human and Machine in-the-Loop Evaluation and Learning Strategies (HAMLETS) workshop at NeurIPS 2020.

- Email: max@bartolo.ai
- Web page: https://www.maxbartolo.com/
- Google Scholar: https://scholar.google.co.uk/citations?user=jPSWYn4AAAAJ

Oana Inel is a Postdoctoral Researcher at the University of Zurich. Her research focuses on measuring the quality of human-annotated and human-generated data and investigating the use of explanations to support people in decision-making. Previously, she was a Postdoctoral Researcher at TU Delft and she received her PhD at the Vrije Universiteit Amsterdam, where her research focused on detecting and representing events and their semantics for understanding knowledge on the web. She has co-organised workshops and tutorials in the area of human computation, explanations for decision-support systems, semantic web technologies at TheWebConf, UMAP, ISWC, and SIGIR.

- Email: inel@ifi.uzh.ch
- Web page: https://oana-inel.github.io
- Google Scholar: https://scholar.google.com/citations?user=mEi2gvgAAAAJ

Rafael Mosquera is a machine learning engineer at MLCommons, where he specializes in developing benchmarks for different ML tasks, as well as the creation of new datasets. He holds a Bachelor's degree in Economics and Law and is currently pursuing a Master's degree in Economics. Rafael has extensive experience in creating open-source datasets for commercial usage and has previously worked on projects such as The People's Speech and Dollar Street. Currently, he leads the implementation of the DataPerf suite of challenges as one of Dynabench main developers.

- Email: rafael.mosquera@mlcommons.org
- Web page: https://www.linkedin.com/in/rafael-mosquera/
- Google Scholar: https://scholar.google.com/citations?user=XC9DJhUAAAAJ

Vijay Janapa Reddi is an Associate Professor at Harvard University, as well as a founding member and Vice President of MLCommons (mlcommons.org), the non-profit organization responsible for hosting the Adversarial Nibbler challenge. With respect to this challenge, his expertise and contribution is primarily focused on constructing robust ML benchmarks that scale. His experience stems from several of the MLCommons benchmarks he developed, including those used in the DataPerf suite, to which the Adversarial Nibbler challenge belongs. Additionally, he has coordinated over 30 workshops and tutorials, and played a significant role in establishing the diverse community surrounding MLCommons by bringing together experts from various fields, which is beneficial for this challenge. Dr. Reddi earned his Ph.D. in Computer Science from Harvard University.

- Email: vj@eecs.harvard.edu
- Web page: http://scholar.harvard.edu/vijay-janapa-reddi/
- Google Scholar: https://scholar.google.com/citations?hl=en&user=gy4UVGcAAAAJ

Will Cukierski is the Head of Competitions at Kaggle. He received his PhD in Biomedical Engineering from Rutgers University in 2012, focusing on applications of ML within cancer diagnostics and imaging. He has served as chair of the KDD Cup and organized hundreds of ML challenges over the last decade.

- Email: wjc@google.com
- Google Scholar: https://scholar.google.com/citations?user=btZpioYAAAAJ

References

Lora Aroyo and Praveen Paritosh. Uncovering unknown unknowns in machine learning. URL https://ai.googleblog.com/2021/02/uncovering-unknown-unknowns-in-machine.html.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. Beat the AI: Investigating adversarial human annotation for reading comprehension. Transactions of the Association for Computational Linguistics, 8:662–678, 2020.

Max Bartolo, Tristan Thrush, Sebastian Riedel, Pontus Stenetorp, Robin Jia, and Douwe Kiela. Models in the loop: Aiding crowdworkers with generative annotation assistants. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3754–3767, 2022.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: Misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963, 2021.

Samuel Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4843–4855, 2021.

Cats4ML. Cats4ML challenge. URL https://cats4ml.humancomputation.com/.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. arXiv preprint arXiv:2202.04053, 2022.

Kate Crawford and Trevor Paglen. Excavating AI: The politics of images in machine learning training sets. AI & SOCIETY, pages 1–12, 2021.

DeepLearning.AI. Data-centric AI competition. URL https://https-deeplearning-ai.github.io/data-centric-comp/.

Leon Derczynski, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, MR Leiser, and Saif Mohammad. Assessing language model deployment with risk cards. arXiv preprint arXiv:2303.18190, 2023.

Hayden Field. How Microsoft and Google use AI red teams to "stress test" their systems, 2022. URL https://www.emergingtechbrew.com/stories/2022/06/14/how-microsoft-and-google-use-ai-red-teams-to-stress-test-their-system. Accessed on 03/08/23.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv preprint arXiv:2209.07858, 2022.

Naman Goel and Boi Faltings. Crowdsourcing with fairness, diversity and budget constraints. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pages 297–304, 2019.

Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445423. URL https://doi.org/10.1145/3411764.3445423.

Kohta Katsuno, Masaki Matsubara, Chiemi Watanabe, and Atsuyuki Morishima. Improving reproducibility of crowdsourcing experiments. 2019.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in NLP. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4110–4124, 2021.

Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. Handling and presenting harmful text in NLP research. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 497–510, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.35.

Hannah Kirk, Bertie Vidgen, Paul Röttger, Tristan Thrush, and Scott Hale. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1352–1368, 2022b.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of BERT. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4365–4374, 2019.

Alexandra Sasha Luccioni and Joseph D Viviano. What's in the box? a preliminary analysis of undesirable content in the common crawl corpus. arXiv preprint arXiv:2105.02732, 2021.

Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Douwe Kiela, David Jurado, et al. Dataperf: Benchmarks for data-centric AI development. arXiv preprint arXiv:2207.10062, 2022.

Midjourney. Midjourney documentation and user guide. https://docs.midjourney.com/, 2023. (Accessed on 04/19/2023).

Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Auditing large language models: A three-layered approach. arXiv preprint arXiv:2302.08500, 2023.

Madhumita Murgia. OpenAI's red team: the experts hired to 'break' ChatGPT. URL https://www.ft.com/content/0876687a-f8b7-4b39-b513-5fee942831e8.

OpenAI. DALL·E 2 pre-training mitigations. https://openai.com/research/dall-e-2-pre-training-mitigations, June 2022. (Accessed on 04/25/2023).

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. DynaSent: A dynamic benchmark for sentiment analysis. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2388–2404, 2021.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pages 33–44, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents, 2022.