CS 269: Computational Ethics, Large Language Models and the Future of NLP

Assignment 1

1. Carlini, Nicholas, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Eric Wallace, David Rolnick and Florian Tramèr. “Stealing Part of a Production Language Model.” 41st International Conference on Machine Learning (2024). https://arxiv.org/abs/2403.06634

2. Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What Does it Mean for a Language Model to Preserve Privacy? In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 2280–2292. https://doi.org/10.1145/3531146.3534642

3. Mireshghallah, Niloofar, Maria Antoniak, Yash More, Yejin Choi and G. Farnadi. “Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild.” Proceedings of the 2024 Conference On Language Modeling (COLM). https://maria-antoniak.github.io/resources/2024_colm_trust_no_bot.pdf

4. Cooper, A. Feder, Christopher A. Choquette-Choo, Miranda Bogen, Matthew Jagielski, Katja Filippova, Ken Ziyu Liu, Alexandra Chouldechova, Jamie Hayes, Yangsibo Huang, Niloofar Mireshghallah, Ilia Shumailov, Eleni Triantafillou, Peter Kairouz, Nicole Mitchell, Percy Liang, Daniel E. Ho, Yejin Choi, Sanmi Koyejo, Fernando Delgado, James Grimmelmann, Vitaly Shmatikov, Christopher De Sa, Solon Barocas, Amy Cyphert, Mark A. Lemley, danah boyd, Jennifer Wortman Vaughan, Miles Brundage, David Bau, Seth Neel, Abigail Z. Jacobs, Andreas Terzis, Hanna Wallach, Nicolas Papernot and Katherine Lee. “Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice.” (2024). https://arxiv.org/abs/2412.06966

Assignment 2

1. Bommasani, Rishi, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky and Percy Liang. “Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization?” Proceedings of Neural Information Processing Systems (2022). https://www.semanticscholar.org/reader/f20d68e2f54d5534746a4fba1c7b01895823e769

2. Bui, M. L. and Noble, S.U. (2020). We're Missing a Moral Framework of Justice in Artificial Intelligence: On the Limits, Failings, and Ethics of Fairness. Oxford University Press Handbook of Ethics of AI. Markus Dubber, Frank Pasquale, and Sunit Das (Eds). Oxford University Press. Oxford, UK. Shared on Perusall.

3. Shomik Jain, Vinith Suriyakumar, Kathleen Creel, and Ashia Wilson. 2024. Algorithmic Pluralism: A Structural Approach To Equal Opportunity. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). Association for Computing Machinery, New York, NY, USA, 197–206. https://doi.org/10.1145/3630106.3658899

4. Fleisig, Eve, Su Lin Blodgett, Dan Klein and Zeerak Talat. "The Perspectivist Paradigm Shift: Assumptions and Challenges of Capturing Human Labels." North American Chapter of the Association for Computational Linguistics (2024). https://www.semanticscholar.org/reader/d341da8368e950c0efc44d74ba3c33c9c5f1314e

Assignment 3

1. Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. 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 (FAT* '20). Association for Computing Machinery, New York, NY, USA, 33–44. https://doi.org/10.1145/3351095.3372873

2. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical Machine Learning in Healthcare. Annu Rev Biomed Data Sci. 2021 Jul;4:123-144. doi: 10.1146/annurev-biodatasci-092820-114757. Epub 2021 May 6. PMID: 34396058; PMCID: PMC8362902. https://pmc.ncbi.nlm.nih.gov/articles/PMC8362902/

3. Umeton, Renato, Anne Kwok, Rahul Maurya, Domenic Leco, Naomi Lenane, Jennifer Willcox, Gregory A. Abel, Mary Tolikas and Jason M. Johnson. "GPT-4 in a Cancer Center — Institute-Wide Deployment Challenges and Lessons Learned." NEJM AI (2024). https://osf.io/preprints/osf/bqv4f

4. Omiye, J.A., Lester, J.C., Spichak, S. et al. Large language models propagate race-based medicine. npj Digit. Med. 6, 195 (2023). https://doi.org/10.1038/s41746-023-00939-z

Assignment 4

1. Jeffrey T Hancock, Mor Naaman, Karen Levy, AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations, Journal of Computer-Mediated Communication, Volume 25, Issue 1, January 2020, Pages 89–100, https://doi.org/10.1093/jcmc/zmz022

2. H. Hosseinmardi, A. Ghasemian, M. Rivera-Lanas, M. Horta Ribeiro, R. West, D.J. Watts, Causally estimating the effect of YouTube's recommender system using counterfactual bots, Proc. Natl. Acad. Sci. U.S.A. 121 (8) e2313377121, https://doi.org/10.1073/pnas.2313377121 (2024).

3. Chen, Canyu and Kai Shu. "Can LLM-Generated Misinformation Be Detected?" International Conference on Learning Representations (2023). https://www.semanticscholar.org/reader/6f75e8b61f13562237851d8119cb2f9d49e073fb

4. Augenstein, I., Baldwin, T., Cha, M. et al. Factuality challenges in the era of large language models and opportunities for fact-checking. Nat Mach Intell 6, 852–863 (2024). https://doi.org/10.1038/s42256-024-00881-z

Assignment 5

1. Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

2. Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. 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 3407–3412, Hong Kong, China. Association for Computational Linguistics. https://aclanthology.org/D19-1339/

3. Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3356–3369, Online. Association for Computational Linguistics. https://aclanthology.org/2020.findings-emnlp.301/

4. Hofmann, V., Kalluri, P.R., Jurafsky, D. et al. AI generates covertly racist decisions about people based on their dialect. Nature 633, 147–154 (2024). https://doi.org/10.1038/s41586-024-07856-5

Assignment 6

1. Longpre, Shayne, Robert Mahari, Naana Obeng-Marnu, William Brannon, Tobin South, Katy Gero, Sandy Pentland and Jad Kabbara. “Data Authenticity, Consent, & Provenance for AI are all broken: what will it take to fix them?” International Conference on Machine Learning (2024). https://www.semanticscholar.org/reader/a15e462e30a1785f53e5eb9210acdf9908e8d7cb

2. Bowman, Sam, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, John Kernion, Jamie Kerr, Jared Mueller, Jeff Ladish, Joshua D. Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noem'i Mercado, Nova Dassarma, Robin Larson, Sam McCandlish, Sandip Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom B. Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Benjamin Mann and Jared Kaplan. “Measuring Progress on Scalable Oversight for Large Language Models.” ArXiv abs/2211.03540 (2022). https://www.semanticscholar.org/reader/99ca5162211a895a5dfbff9d7e36e21e09ca646e

3. Gebru, T., & Torres, Émile P. (2024). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. First Monday, 29(4). https://doi.org/10.5210/fm.v29i4.13636

4. Landau, Susan, James X. Dempsey, Ece Kamar, Steven M. Bellovin and Robert Pool. “Challenging the Machine: Contestability in Government AI Systems.” ArXiv abs/2406.10430 (2024). https://arxiv.org/pdf/2406.10430