

CS 4390/5390, Spring 2027

Human-Related Performance Metrics for AI Systems

Instructor: Nigel Ward nigel@utep.edu

Mondays and Wednesdays, 15:00 – 16:00, CCSB room 1.0202

Motivation: AI systems often mimic, interact with, assess, or support humans. Human-subjects experiments are the gold standard, but are costly, so there is a need for loss functions, performance metrics, and other proxies to support the training, alignment, and evaluation of such systems. This course will examine ways to create such loss functions and metrics. We will cover the basic issues in metric design and explore strategies and methods through a broad sampling of case studies, covering such topics as visual and speech perception, decision-making, responsiveness, evaluation of interactive systems, skill assessment, perceptions of privacy, multi-objective optimization, and questions of individual and societal value.

Format: This course will be primarily discussion based: we will read and discuss one or two substantial papers every week. Readings will cover foundations and topics in metrics and adjacent areas, depending on the interests of the participants. There will be a final report and presentation on the application of these ideas to a specific problem or topic, ideally in the student's area of interest.

Audience: Knowledge of basic statistical methods or machine learning, experiment design, human perception, human cognition, requirements analysis, design of interactive systems, usability engineering and other topics in human-computer interaction, evaluation of models, design of benchmarks. The target audience is students of computer science, engineering, and psychology, but others are also welcome.

Readings may include some of the following:

Ara, Jinat, and Cecilia Sik-Lanyi. "Automated evaluation of accessibility issues of webpage content: tool and evaluation." *Scientific Reports* 15, no. 1 (2025): 9516.

Cheng, Myra, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. "Elephant: Measuring and understanding social sycophancy in LLMs." In *International conference on learning representations*, vol. 2026, pp. 130060-130097. 2026.

- Frese, Thomas, Charles A. Bouman, and Jan P. Allebach. "Methodology for designing image similarity metrics based on human visual system models." In *Human Vision and Electronic Imaging II*, vol. 3016, pp. 472-483. SPIE, 1997.
- Goyal, Anirudh, and Yoshua Bengio. "Inductive biases for deep learning of higher-level cognition." *Proceedings of the Royal Society A* 478, no. 2266 (2022): 20210068.
- Ibrahim, Lujain, Saffron Huang, Lama Ahmad, Umang Bhatt, and Markus Anderljung. "Towards interactive evaluations for interaction harms in human-AI systems." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 8, no. 2, pp. 1302-1310. 2025.
- Jurafsky, Daniel S. and James H. Martin. *Post-training. Sections 8.3 and 8.4 of Speech and Language Processing*, 2026.
- Khan, Aftab, Sebastian Mellor, Eugen Berlin, Robin Thompson, Roisin McNaney, Patrick Olivier, and Thomas Plötz. "Beyond activity recognition: skill assessment from accelerometer data." In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pp. 1155-1166. 2015.
- Lee, Min Hun. "From Accuracy to Readiness: Metrics and Benchmarks for Human-AI Decision-Making" In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1-10. 2026.
- Milli, Smitha, Luca Belli, and Moritz Hardt. "From optimizing engagement to measuring value." In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 714-722. 2021.
- Papi, Sara, Peter Polak, Dominik Macháček, and Ondřej Bojar. "How "real" is your real-time simultaneous speech-to-text translation system?" *Transactions of the Association for Computational Linguistics* 13 (2025): 281-313.
- Ren, Yaxuan, Krithika Ramesh, Yaxing Yao, and Anjalie Field. "How do we measure privacy in text? A survey of text anonymization metrics." In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 1532-1544. 2025.
- Ryu, Hansol X., and Manoj Srinivasan. "What loss functions do humans optimize when they perform regression and classification?" *Biorxiv* (2023).
- Shen, Hua, Tiffany Kneare, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Liu Yachuan, Savvas Petridis et al. "Position: Towards bidirectional human-AI alignment." *Advances in Neural Information Processing Systems* 38 (2026).
- Shermis, Mark D., and Joshua Wilson. "Introduction to automated essay evaluation." *The Routledge international handbook of automated essay evaluation* (2024): 3-22.
- Torrance, George W., and David Feeny. "Utilities and quality-adjusted life years." *International journal of technology assessment in health care* 5, no. 4 (1989): 559-575.

- Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. "GLUE: A multi-task benchmark and analysis platform for natural language understanding." In Proceedings of the EMNLP workshop: "BlackboxNLP: Analyzing and interpreting neural networks for NLP", pp. 353-355. 2018.
- Wang, Shihua, Andrew Sekey, and Allen Gersho. "An objective measure for predicting subjective quality of speech coders." IEEE Journal on selected areas in communications 10, no. 5 (1992): 819-829.
- Ward, Nigel G., Divette Marco, and Olac Fuentes. "Which prosodic features matter most for pragmatics?" In International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025.
- Yan, Bi-Cheng, Hsin-Wei Wang, Yi-Cheng Wang, Jiun-Ting Li, Chi-Han Lin, and Berlin Chen. "Preserving phonemic distinctions for ordinal regression: A novel loss function for automatic pronunciation assessment." In IEEE automatic speech recognition and understanding workshop (ASRU), 2023.
- Yang, Yifan, Bing Han, Hui Wang, Long Zhou, Wei Wang, Mingyu Cui, Xu Tan, and Xie Chen. "Measuring prosody diversity in zero-shot TTS: A new metric, benchmark, and exploration." In International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 17252-17256. IEEE, 2026.