(http://stanford.edu/)

CS329H: Machine Learning from Human Preferences

Autumn 2023

Content

Machine learning from human preferences provides mechanisms for capturing human feedback, which is used to design reward functions that are otherwise difficult to specify quantitatively, e.g., for socio-technical applications such as algorithmic fairness and many language and robotic tasks. While learning from human preferences has emerged as an increasingly important component of modern machine learning, e.g., credited with advancing the state of the art in language modeling and reinforcement learning, existing approaches are largely reinvented independently in each subfield, with limited connections drawn among them.

This course will cover the foundations of learning from human preferences from first principles and outline connections to the growing literature on the topic. This includes but is not limited to:

- Inverse reinforcement learning, which uses human preferences to specify the reinforcement learning reward function
- Metric elicitation, which uses human preferences to specify tradeoffs for cost-sensitive classification
- Reinforcement learning from human feedback, where human preferences are used to align a pre-trained language model

This is a graduate-level course. By the end of the course, students should be able to understand and implement state-of-the-art learning from human feedback and be ready to research these topics. Given how fast this area is growing, this course will consist of weekly lectures, presentations, and discussions of papers led by students. Students will compile course notes along with a final course project. If you are a CS PhD student at Stanford, this course is counted toward the breath requirement (https://www.cs.stanford.edu/phd-program-foundation-and-breadth-requirements) for "Learning and Modeling" or "Human and Society".

Instructor

Sanmi Koyejo
(https://cs.stanford.edu/~sanmi/)

Course Assistant

Sang Truong
(https://ai.stanford.edu/~sttruong)

Logistics

- **Lectures** are on Monday and Wednesday from 1:30 PM - 2:50 PM PT in building 370 room 370. There is no remote option and the lectures will not be recorded.
- **Contact**: Post on Ed (https://edstem.org/us/courses/48383/discussion/) for any questions. For private matters that can not be handled via Ed, email sttruong [AT] cs [DOT] stanford [DOT] edu cc' sanmi [AT] cs [DOT] stanford [DOT] edu.
- **Office Hours**:
  - Sanmi: 10:00 AM - 11:00 AM on Monday in Gates 332.
  - Sang: 1:00 PM - 3:00 PM on Friday by appointment (https://calendar.app.google/kffvM45XmZHXFVhY8) on Zoom (https://stanford.zoom.us/j/7657210999?pwd=b3V0ZHY4ZWh6L3FOdjZpNDE1M2owUT09) (by default) or Gates 234.
- **Auditor Policy**: We welcome auditors if there is enough classroom space. Auditors are asked to be peer-reviewers.

Schedule

The current class schedule is below (subject to change). A tentative reading list can be found here (https://docs.google.com/document/d/1gGzR442HO0ebh_nW4m5rx7j2um8Y6xYgF4g6mFNLBDY/edit?usp=sharing).

Date: **Week 1: Sep 27**
Description: **[Lecture] Course Introduction. (https://web.stanford.edu/class/cs329h/slides/cs329H-lec1.pdf)**
Recommended reading: None
Events: Deadline: 1. Sign-up for Presentation and Scribe (https://docs.google.com/spreadsheets/d/1a7p6gGa6FWGLy8IUtIOD_QRUIwK2Cf88Bz4D92DanbA/edit)

Date: **Week 2: Oct 2**
Description: **[Lecture] Human preferences models. (https://web.stanford.edu/class/cs329h/slides/cs329H-lec2.pdf)**
Recommended reading: 1. Train. Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand (https://mitpress.mit.edu/9780262519465/qualitative-choice-analysis/). MIT Press. 1985. 2. McFadden, Train. Mixed MNL Models for Discrete Response (https://onlinelibrary.wiley.com/doi/abs/10.1002/1099-1255%28200009/10%2915%3A5%3C447%3A%3AAID-JAE570%3E3.0.CO%3B2-1#). Journal of Applied Econometrics. 2000. 3. Luce. Individual Choice Behavior: A Theoretical Analysis (https://psycnet.apa.org/record/1960-03588-000). Wiley. 1959. Additional reading: 1. Ben-Akiva, Lerman. Discrete Choice Analysis: Theory and Application to Travel Demand (https://mitpress.mit.edu/9780262536400/discrete-choice-analysis/). Transportation Studies. 1985. 2. Park, Simar, Zelenyuk. Nonparametric Estimation of Dynamic Discrete Choice Models for Time Series Data (https://www.sciencedirect.com/science/article/pii/S0167947316302596). Computational Statistics & Data Analysis. 2017. 3. Rafailov, Sharma, Mitchell, Ermon, Manning, Finn. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model (https://arxiv.org/abs/2305.18290). Preprint. 2023.
Events: Deadline: 1. Pre-class survey (https://forms.gle/Ztj97pKdwWuBwm7m6)

Date: **Week 2: Oct 4**

Description: **[Student Presentation] Interaction models (https://web.stanford.edu/class/cs329h/slides/cs329h-lec3.pdf)**

Recommended reading:

1. Cattelan. Models for Paired Comparison Data: A Review with Emphasis on Dependent Data (https://arxiv.org/abs/1210.1016). Statistical Science. 2012.
2. Bhatia, Pananjady, Bartlett, Dragan, Wainwright. Preference Learning Along Multiple Criteria: A Game-Theoretic Perspective (https://proceedings.neurips.cc/paper/2020/hash/52f4691a4de70b3c441bca6c546979d9-Abstract.html). NeurIPS. 2020.
3. Shah, Gundotra, Abbeel, Dragan. On the Feasibility of Learning, Rather Than Assuming, Human Biases for Reward Inference (https://arxiv.org/abs/1906.09624). ICML. 2019.
4. Ghosal, Zurek, Brown, Dragan. The Effect of Modeling Human Rationality Level on Learning Rewards from Multiple Feedback Types (https://ojs.aaai.org/index.php/AAAI/article/view/25740). AAAI. 2023.

Events: Deadline:

1. Presentation slide and Presentation feedback for "Interaction models".

Date: **Week 3: Oct 9**

Description: **[Fireside chat] Psychology and Marketing Perspectives: Noah Goodman, Jonathan Levav, S. Christian Wheeler**

Additional reading:

1. Evangelidis, Levav, Simonson. The Upscaling Effect: How the Decision Context Influences Tradeoffs Between Desirability and Feasibility (https://academic.oup.com/jcr/article/50/3/492/6935792). Journal of Consumer Research. 2023.
2. Evangelidis, Levav, Simonson. A Reexamination of the Impact of Decision Conflict on Choice Deferral (https://pubsonline.informs.org/doi/full/10.1287/mnsc.2022.4484). Management Science. 2023.
3. Shennib, Catapano, Levav. Preference Reversals Between Digital and Physical Goods (https://journals.sagepub.com/doi/full/10.1177/00222437211065020). ACR North American Advances. 2019.
4. Tamkin, Handa, Shrestha, Goodman. Task Ambiguity in Humans and Language Models (https://arxiv.org/abs/2212.10711). arXiv. 2022.
5. Hawkins, Berdahl, Pentland, Tenenbaum, Goodman, Krafft. Flexible Social Inference Facilitates Targeted Social Learning When Rewards Are Not Observable (https://arxiv.org/abs/2212.00869). Nature Human Behaviour. 2023.
6. Yu, Goodman, Mu. Characterizing Tradeoffs Between Teaching via Language and Demonstrations in Multi-Agent Systems (https://arxiv.org/abs/2305.11374). arXiv. 2023.

Events: Deadline: None

Date: **Week 3: Oct 11**

Description: **[Student Presentation] Human biases and Reward models (https://web.stanford.edu/class/cs329h/slides/cs329h-lec5.pdf)**

Recommended reading:

1. The Decision Lab. Biases Index (https://thedecisionlab.com/biases-index). 2023.
2. Slovic. The Construction of Preference (https://doi.org/10.1017/CBO9780511618031). Shaping Entrepreneurship Research. 2020.
3. Hogarth. Insights in Decision Making: A Tribute to Hillel J. Einhorn (https://press.uchicago.edu/ucp/books/book/chicago/I/bo3638587.html). University of Chicago Press. 1990.
4. Cooke. Experts in Uncertainty: Opinion and Subjective Probability in Science (https://psycnet.apa.org/record/1991-98990-000). Oxford University Press. 1991.
5. Chan, Critch, Dragan. Human Irrationality: Both Bad and Good for Reward Inference (https://arxiv.org/abs/2111.06956). arXiv. 2021.
6. Bobu, Scobee, Fisac, Sastry, Dragan. Less is More: Rethinking Probabilistic Models of Human Behavior (https://dl.acm.org/doi/abs/10.1145/3319502.3374811). ACM/IEEE International Conference on Human-Robot Interaction. 2020.

Events: Deadline:

1. Presentation slide and Presentation feedback for "Human biases and Reward models"

Date: **Week 4: Oct 16**

Description: **[Student Presentation] Metric elicitation (https://web.stanford.edu/class/cs329h/slides/cs329h-lec6.pdf)**

Recommended reading:

1. Hiranandani, Boodaghians, Mehta, Koyejo. Performance Metric Elicitation from Pairwise Classifier Comparisons (https://proceedings.mlr.press/v89/hiranandani19a.html). AISTATS. 2019.
2. Hiranandani, Boodaghians, Mehta, Koyejo. Multiclass Performance Metric Elicitation (https://papers.nips.cc/paper_files/paper/2019/hash/1fd09c5f59a8ff35d499c0ee25a1d47e-Abstract.html). NeurIPS. 2019.
3. Hiranandani, Narasimhan, Koyejo. Fair Performance Metric Elicitation (https://proceedings.neurips.cc/paper/2020/hash/7ec2442aa04c157590b2fa1a7d093a33-Abstract.html). NeurIPS. 2020.
4. Hiranandani, Mathur, Narasimhan, Koyejo. Quadratic Metric Elicitation with Application to Fairness (https://proceedings.mlr.press/v180/hiranandani22a.html). UAI. 2022.

Additional reading:

1. Ali, Upadhyay, Hiranandani, Glassman, Koyejo. Metric Elicitation: Moving from Theory to Practice (https://arxiv.org/abs/2212.03495). NeurIPS Workshop on Human-Centered AI (HCAI), 2022.
2. Riabacke, Danielson, Ekenberg, L. State-of-the-Art Prescriptive Criteria Weight Elicitation (https://www.hindawi.com/journals/ads/2012/276584/). Advances in Decision Sciences, 2012.

Events: Deadline:

1. Presentation slide and Presentation feedback for "Metric elicitation"
2. Scribe for "Human preferences models"

Date: **Week 4: Oct 18**

Description: **[Student Presentation] Active learning (https://web.stanford.edu/class/cs329h/slides/cs329h-lec7.pdf)**

Recommended Readings:

1. Cohn, Ghahramani, Jordan. Active Learning with Statistical Models (https://www.jair.org/index.php/jair/article/view/10158). JAIR. 1996.
2. Biyik, Sadigh. Batch Active Preference-Based Learning of Reward Functions (https://proceedings.mlr.press/v87/biyik18a.html). CORL. 2018.
3. Sadigh, Dragan, Sastry, Seshia. Active Preference-Based Learning of Reward Functions (https://escholarship.org/uc/item/88k894w7). UC Berkeley. 2017.
4. Jamieson, Nowak. Active Ranking Using Pairwise Comparisons (https://proceedings.neurips.cc/paper/2011/hash/6c14da109e294d1e8155be8aa4b1ce8e-Abstract.html). NeurIPS. 2011.
5. Holladay, Javdani, Dragan, Srinivasa. Active Comparison Based Learning Incorporating User Uncertainty and Noise (https://www.researchgate.net/profile/Rachel-Holladay/publication/357875816_Active_Comparison_Based_Learning_Incorporating_User_Uncertainty_and_Noise/link Comparison-Based-Learning-Incorporating-User-Uncertainty-and-Noise.pdf). RSS Workshop on Model Learning for Human-Robot Communication. 2016.

Additional Readings:

1. Settles. Active Learning Literature Survey (http://digital.library.wisc.edu/1793/60660). University of Wisconsin-Madison. 2009.

Events: Deadline:

1. Presentation slide and Presentation feedback for "Active learning".
2. Scribe for "Interaction models".

Date: **Week 5: Oct 23**

Description: **[Student Presentation] Bandits and Probabilistic Methods (https://web.stanford.edu/class/cs329h/slides/cs329h-lec8.pdf)**

Recommended Readings:

1. Agarwal, Hsu, Kale, Langford, Li, Schapire. Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits (https://proceedings.mlr.press/v32/agarwalb14.html). In International Conference on Machine Learning, pp. 1638-1646. PMLR, 2014.
2. Bouneffouf, Rish, Aggarwal. Survey on Applications of Multi-Armed and Contextual Bandits. In 2020 IEEE Congress on Evolutionary Computation (CEC), pp. 1-8. IEEE, 2020.
3. Sui, Zoghi, Hofmann, Yue. Advancements in Dueling Bandits (https://www.ijcai.org/proceedings/2018/0776.pdf). In IJCAI, pp. 5502-5510. 2018.
4. Yue, Broder, Kleinberg, Joachims. The K-Armed Dueling Bandits Problem (https://www.sciencedirect.com/science/article/pii/S0022000012000281). Journal of Computer and System Sciences 78, no. 5 (2012): 1538-1556.

Events: Deadline:

1. Proposal deadline
2. Presentation slide and Presentation feedback for "Bandits and Probabilistic Methods".
3. Scribe feedback for "Human preferences models"

Date: **Week 5: Oct 25**

Description: **[Students Presentation] Multimodal rewards; Meta reward learning (https://web.stanford.edu/class/cs329h/slides/cs329h-lec9.pdf)**

Recommended reading:

1. Hejna, Sadigh. Few-Shot Preference Learning for Human-in-the-Loop RL (https://openreview.net/forum?id=IKC5TfXLuW0). CoRL, 2023.
2. Zhou, Jang, Kappler, Herzog, Khansari, Wohlhart, Bai, Kalakrishnan, Levine, Finn. Watch, Try, Learn: Meta-Learning from Demonstrations and Reward (https://arxiv.org/abs/1906.03352). Arxiv, 2019.
3. Myers, Bıyık, Anari, Sadigh. Learning Multimodal Rewards from Rankings (https://arxiv.org/abs/2109.12750). Arxiv, 2021.

Events: Deadline:

1. Presentation slide for "Multimodal rewards; Meta reward learning"
2. Scribe for "Human biases and Reward models"
3. Scribe feedback for "Interaction models"

Date: **Week 6: Oct 30**

Description: **[Guest lecture] Pat Langley (http://www.isle.org/langley/) (Institute for the Study of Learning and Expertise): Human computing (https://web.stanford.edu/class/cs329h/slides/cs329h-lec10.pdf)**

Recommended reading:

Events: Deadline:

1. Scribe for "Metric elicitation"
2. Scribe rebuttal for "Human preferences models"

Date: **Week 6: Nov 1**

Description: **[Student presentation] Alignment; Expert and non-expert stakeholders (https://web.stanford.edu/class/cs329h/slides/cs329h-lec11.pdf)**

Recommended reading:

1. Brown, Schneider, Dragan, Due Niekum. Value Alignment Verification (http://proceedings.mlr.press/v139/brown21a.html). ICML. 2021.
2. Bobu, Bajcsy, Fisac, Due Dragan. Learning under Misspecified Objective Spaces (https://proceedings.mlr.press/v87/bobu18a). CoRL. 2018.
3. Jeon, Milli, Due Dragan. Reward-Rational (Implicit) Choice: A Unifying Formalism for Reward Learning (https://proceedings.neurips.cc/paper/2020/hash/2f10c1578a0706e06b6d7db6f0b4a6af-Abstract.html). NeurIPS. 2020.
4. Bobu, Peng, Agrawal, Shah, Due Dragan. Aligning Robot and Human Representations (https://arxiv.org/abs/2302.01928). arXiv. 2023.

Events: Deadline

1. Presentation slide and Presentation feedback for "Alignment; Expert and non-expert stakeholders"
2. Scribe for "Active learning"
3. Scirbe feedback for "Human biases and Reward models"
4. Scribe rebuttal for "Interaction models"

Date: **Week 7: Nov 6**

Description: **[Guest lecture] Meredith Ringel Morris (https://cs.stanford.edu/~merrie/) (Google DeepMind): HCI considerations in learning from humans (Virtual) (https://web.stanford.edu/class/cs329h/slides/cs329h-lec11.pdf)**

Recommended reading:

Events: Deadline

1. Scribe for Pat Langley
2. Scribe for "Bandits and Probabilistic Methods"
3. Scribe feedback for "Metric elicitation"

Date: **Week 7: Nov 8**

Description: **[Guest lecture] Vasilis Syrgkanis (https://vsyrgkanis.com/)(Stanford): Truthfulness and mechanism design (https://web.stanford.edu/class/cs329h/slides/cs329h-lec12.pdf)**

Recommended reading:

1. Balcan, Sandholm, VitercikTutorial on Mechanism Design (https://sites.google.com/view/amdtutorial/home). 2023
2. Roughgarden Lectures 1 & 2 on the General Mechanism Design Problem and the Idea of Incentive Compatibility (https://timroughgarden.org/f13/l/).
3. Linstone, Turoff. The Delphi Method. Reading, MA: Addison-Wesley, 1975.
4. Prelec. A Bayesian Truth Serum for Subjective Data (https://www.science.org/doi/10.1126/science.1102081). Science. 2004.

Events: Deadline

1. Scribe for "Multimodal rewards; Meta reward learning"
2. Scribe feedback for "Active learning"
3. Scirbe rebuttal for "Human biases and Reward models"

Date: **Week 8: Nov 13**

Description: **[Guest lecture] Jason Hartline (https://sites.northwestern.edu/hartline/) (Northwestern): Truthfulness and mechanism design (https://web.stanford.edu/class/cs329h/slides/cs329h-lec13.pdf)**

Recommended reading:

1. Schenk, Guittard. Crowdsourcing: What Can Be Outsourced to the Crowd, and Why? (https://inria.hal.science/halshs-00439256v1). HAL Open Science. 2009.
2. Quinn, Bederson. Human Computation: A Survey and Taxonomy of a Growing Field (http://alexquinn.org/papers/Human%20Computation,%20A%20Survey%20and%20Taxonomy%20of%20a%20Growing SIGCHI Conference on Human Factors in Computing Systems. 2011.
3. Kong. Dominantly Truthful Multi-Task Peer Prediction with a Constant Number of Tasks (https://arxiv.org/abs/1911.00272). ACM-SIAM Symposium on Discrete Algorithms. 2020.
4. Kong, Schoenebeck. An Information Theoretic Framework for Designing Information Elicitation Mechanisms That Reward Truth-Telling (https://dl.acm.org/doi/abs/10.1145/3296670). ACM Transactions on Economics and Computation. 2019.

Events: Deadline

1. Scribe for Meredith Ringel Morris
2. Scribe feedback for Pat Langley
3. Scribe feedback for "Bandits and Probabilistic Methods"
4. Scribe rebuttal for "Metric elicitation"

Date: **Week 8: Nov 15**

Description: **[Guest lecture] Dorsa Sadigh (https://dorsa.fyi/) (Stanford): Inverse reinforcement learning from human feedback for robotics (https://web.stanford.edu/class/cs329h/slides/cs329h-lec14.pdf)**

Recommended reading:

1. Ng, Russell. Algorithms for Inverse Reinforcement Learning (https://ai.stanford.edu/~ang/papers/icml00-irl.pdf). ICML. 2000.
2. Hadfield-Menell, Russell, Abbeel, Dragan. Cooperative Inverse Reinforcement Learning (https://proceedings.neurips.cc/paper_files/paper/2016/hash/c3395dd46c34fa7fd8d729d8cf88b7a8-Abstract.html). NeurIPS. 2016.
3. Arora, Doshi. A Survey of Inverse Reinforcement Learning: Challenges, Methods and Progress (https://arxiv.org/abs/1806.06877). Artificial Intelligence. 2021.
4. Hadfield-Menell, Milli, Abbeel, Russell, Dragan. Inverse Reward Design (https://proceedings.neurips.cc/paper/2017/hash/32fdab6559cdfa4f167f8c31b9199643-Abstract.html). NeurIPS. 2017.
5. Shin, Dragan, Brown. Benchmarks and Algorithms for Offline Preference-Based Reward Learning (https://arxiv.org/abs/2301.01392). arXiv. 2023.
6. Ghosal, Zurek, Brown, Dragan. The Effect of Modeling Human Rationality Level on Learning Rewards from Multiple Feedback Types (https://ojs.aaai.org/index.php/AAAI/article/view/25740). AAAI. 2023.
7. Bıyık, Losey, Palan, Landolfi, Shevchuk, Sadigh. Learning Reward Functions from Diverse Sources of Human Feedback: Optimally Integrating Demonstrations and Preferences (https://journals.sagepub.com/doi/full/10.1177/02783649211041652). The International Journal of Robotics Research. 2022.

Events: Deadline

1. Scribe for "Alignment; Expert and non-expert stakeholders"
2. Scribe for Vasilis Syrgkanis
3. Scribe feedback for "Multimodal rewards; Meta reward learning"
4. Scribe rebuttal for Pat Langley
5. Scribe rebuttal for "Active learning"

Date: **Week 9: Nov 20**

Description: **Thanksgiving Recess (no classes)**

Events:

Date: **Week 9: Nov 22**

Description: **Thanksgiving Recess (no classes)**

Events:

Date: **Week 10: Nov 27**

Description: **[Guest Lecture] Diyi Yang (https://cs.stanford.edu/~diyiy/) (Stanford): Ethics and HCI (https://web.stanford.edu/class/cs329h/slides/cs329h-lec15.pdf)**

Recommended reading:

1. Busarovs. Ethical Aspects of Crowdsourcing, or Is It a Modern Form of Exploitation (https://ijeba.com/documents/papers/2013_1_p1.pdf). International Journal of Economics & Business Administration. 2013.
2. Denton, Díaz, Kivlichan, Prabhakaran, & Rosen. Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation (https://arxiv.org/abs/2112.04554). arXiv. 2021.

Events: Deadline

1. Project deadline
2. Scribe for Jason Hartline
3. Scribe feedback for Meredith Ringel Morris
4. Scribe rebuttal for "Bandits and Probabilistic Methods"

Date: **Week 10: Nov 29**

Description: **[Guest Lecture] Nathan Lambert (https://www.natolambert.com/) (HuggingFace): Reinforcement learning from human feedback for language models (https://web.stanford.edu/class/cs329h/slides/cs329h-lec16.pdf)**

Recommended Readings:

1. Bansal, Dang, Grover. Peering Through Preferences: Unraveling Feedback Acquisition for Aligning Large Language Models (https://arxiv.org/abs/2308.15812). arXiv. 2023.
2. Christiano, Leike, Brown, Martic, Legg, Amodei. Deep Reinforcement Learning from Human Preferences (https://arxiv.org/abs/1706.03741). NeurIPS. 2017.
3. Ziegler, Stiennon, Wu, Brown, Radford, Amodei, Christiano, Irving. Fine-Tuning Language Models from Human Preferences (https://arxiv.org/abs/1909.08593). arXiv. 2019.

Events: Deadline

1. Scribe for Dorsa Sadigh
2. Scirbe feedback for "Alignment; Expert and non-expert stakeholders"
3. Scribe feedback for Vasilis Syrgkanis
4. Scribe rebuttal for "Multimodal rewards; Meta reward learning"
5. Scribe rebuttal for Meredith Ringel Morris

Date: **Week 11: Dec 4**

Description: **[Lecture] Open Questions & Frontiers (https://web.stanford.edu/class/cs329h/slides/cs329h-lec16.pdf)**

Recommended Readings:

1. Wirth, Akrour, Neumann, Fürnkranz. A Survey of Preference-Based Reinforcement Learning Methods (https://jmlr.org/papers/v18/16-634.html). JMLR, 2017.
2. Casper et al. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback (https://arxiv.org/abs/2307.15217). Arxiv, 2023.

Events: Deadline

1. Scribe for Diyi Yang
2. Scribe feedback for Jason Hartline

Date: **Week 11: Dec 6**
Description: **Poster session**
Recommended Readings: None
Events: Deadline 1. Scribe for Nathan Lambert 2. Scribe feedback for Dorsa Sadigh 3. Scirbe rebuttal for "Alignment; Expert and non-expert stakeholders" 4. Scribe rebuttal for Jason Hartline 5. Scribe rebuttal for Vasilis Syrgkanis

Date: **Week 12: Dec 11**
Description: **Final week: No class**
Events: Deadline 1. Scribe rebuttal for Dorsa Sadigh 2. Scribe feedback for Diyi Yang 3. Scribe feedback for Nathan Lambert

Date: **Week 12: Dec 13**
Description: **Final week: No class**
Events: Deadline 1. Scribe rebuttal for Diyi Yang 2. Scribe rebuttal for Nathan Lambert

Grading

- **Project (50%)**: Complete in groups (max 5 students). If working in a group, include a contribution statement for each deliverable. Aim for a project covering learning from human preferences suitable for submission to a machine learning conference. See the project rubric here (https://docs.google.com/document/d/1-MDLfZ6uBRlS1nDfraKo4jEwUG70HkEb8pOaA_nNtow/edit?usp=sharing).
    - **[Due Oct 23]** Project Proposal (5%), including title and group members, project overview, survey of 3+ relevant papers, description of datasets or simulations (if any), and work division plan. Submit through OpenReview (https://openreview.net/group?id=stanford.edu/Stanford_University/Fall2023/CS329H) in NeurIPS 2023 (https://neurips.cc/Conferences/2023/PaperInformation/StyleFiles) format (pdf+TeX source).
    - **[Due Nov 27]**
        1. Final Manuscript (25%): 8-page max, plus references. Submit through OpenReview (https://openreview.net/group?id=stanford.edu/Stanford_University/Fall2023/CS329H) in NeurIPS 2023 (https://neurips.cc/Conferences/2023/PaperInformation/StyleFiles) format (pdf+TeX source+code+data).
        2. Poster (10%): Submit via OpenReview (https://openreview.net/group?id=stanford.edu/Stanford_University/Fall2023/CS329H).
        3. Medium Post (10%): Submit via OpenReview (https://openreview.net/group?id=stanford.edu/Stanford_University/Fall2023/CS329H).
    - **[Due Dec 06]**
        1. Rebuttal end: Last day to address peer-review concerns.
        2. Medium post deadline: Post on Medium (https://medium.com/) after peer-review and submit link here (https://forms.gle/QZBNHJvTnNj8Pr7G7).
        3. Poster day.

- Presentation and Scribe (40%):
    - [Post-presentation] In-class Presentations (20%): Groups (max 5 students) present on assigned topics. Both peer and instructor evaluations are included in grading. Email presentation to Course Assistant post-class. Presentation guidelines are here (https://docs.google.com/document/d/195iNjLWPTp4RFOG82b_fcVr5a2Ls_v6GxXtomjcOru0/edit?usp=sharing).
    - Scribes (20%): Create 2 scribes (1 from your presentation, 1 from a guest lecture). Aim for textbook-quality content. Submit drafts within 1 or 2 weeks post-presentation on OpenReview (https://openreview.net/group?id=stanford.edu/Stanford_University/Fall2023/CS329H) (pdf+TeX source). Scribes are revised via peer-review. Scribe guidelines are here (https://docs.google.com/document/d/1JqTtXqNqAQyLJVq-UNTzpYcztXDHueCtunTj9Nbu16M/edit?usp=sharing).
        1. [1 or 2 weeks post-presentation] Initial scribe submission.
        2. Receive feedback 2 or 3 weeks post-presentation.
        3. [2.5 or 4 weeks post-presentation] Final revision.
- Peer-review (10%): Complete independently, not in groups.
    - [Due Oct 30] Submit OpenReview ID here (https://forms.gle/5wyMKx8upP1GbEkW8).
    - [Post-presentation] Provide feedback on 2 student presentations (3%). Submit here (https://forms.gle/8FWTqP1ffD3NBUXLA). Review assignments are listed here (https://docs.google.com/spreadsheets/d/1a7p6gGa6FWGLy8IUtIOD_QRUIwK2Cf88Bz4D92DanbA/edit?usp=sharing).
    - [2 or 3 weeks post-presentation/lecture] Review 2 scribes on OpenReview (3%). Ensure your OpenReview ID is submitted on time. Review assignments are here (https://docs.google.com/spreadsheets/d/1a7p6gGa6FWGLy8IUtIOD_QRUIwK2Cf88Bz4D92DanbA/edit?usp=sharing).
    - [Due Dec 04] Feedback for all component of 1 project (4%). Submit on OpenReview.
- Class participation (Extra credit up to 5%): QA on Ed and in-class participation.
- No homework or exam