RESEARCH STATEMENT: EVALUATING AND ENABLING HUMAN–AI COLLABORATION

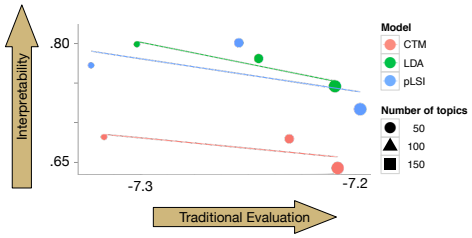
JORDAN BOYD-GRABER, UNIVERSITY OF MARYLAND

Artificial intelligence[1] (AI) is ubiquitous: detecting spam e-mails, flagging fraudulent purchases, and providing the next movie in a Netflix binge. But they do not exist in a vacuum: as Shneiderman [32] argues, AI must exist alongside humans. My goal is to create metrics to measure whether AI methods make sense to users, helping users craft examples to advance AI, and applying AI to illuminate complex social science applications.

1. Evaluating Interpretability

My journey with evaluating interpretability began fifteen years ago with topic models. Topic models are sold as a tool for understanding large data collections: lawyers scouring Nordstream e-mails for a smoking gun, journalists making sense of Wikileaks, or humanists characterizing the oeuvre of Lope de Vega. But topic models' proponents never asked what those lawyers, journalists, or humanists needed. Instead, they optimized *held-out likelihood*.

When my colleagues and I developed the *interpretability* measure to assess whether topic models' users understood their outputs, interpretability and held-out likelihood were negatively correlated [6]! The topic modeling community (including me) had fetishized complexity at the expense of usability. . . and topic modeling is not alone.



Since this humbling discovery, I've built topic models that are a collaboration between humans and computers. The computer starts by proposing an organization of the data. The user separates confusing clusters or joins similar clusters together [18], an improvement over the "take it or leave it"

Date: Updated April 2025.

[1]I take a broad interpretation of AI; some of my examples might be better characterized machine learning. But rather than distracting boundary policing, I will embrace the general term but will be specific in describing particular tools/models.

philosophy of most machine learning algorithms. Focusing on collaboration with as many people as possible also requires algorithms that are low latency [25] and multi-lingual [43].

After we proposed our "reading tea leaves" evaluation, it's heartening that Lau et al. [23] and their "machine reading tea leaves" (which correlate with our human measures) became a standard topic model evaluation: in a survey of forty recent topic modeling papers, **all but four** use a form of their coherence evaluation. However, as we argue in Hoyle et al. [17], you cannot just use this evaluation forever and forget about humans. In that same survey, **none** of those papers do a human evaluation. As topic models evolve (e.g., incorporating neural components), you need to validate that these automatic metrics still correlate with whether it is useful for a human–computer collaboration [24].

2. Teaming as an Evaluation

Within the HCI community, we have argued for the foundations of what should go into human–computer collaborations: computers that incorporate users' suggestions [22]; explanations with accountability [34]; and stable explanations [35].

In addition to these human-centered understanding of users' needs and desires, we've developed machine learning approaches to measure how well users complete a task. For example, for a question answering task, we measured how much the accuracy of the human–computer *team* increases with different explanations and found that explanations help all users but that novices are easily overwhelmed [10]. In follow-on work, we learned how to explicitly optimize explanations for individual users [11] or to help users in negotiations [14].

3. Connecting with Social Science: Pedagogy, Framing, and Deception

The reverse of cooperation is human competition; it also has much to teach computers. I've increasingly looked at language-based games whose clear goals and intrinsic fun speed research progress. For example, in the board game *Diplomacy*, users chat with each other while marshaling armies for world conquest. Alliances are fluid: friends are betrayed and enemies embraced as the game develops. However, users' conversations let us predict when friendships break.

Thus, we argued that Diplomacy would be an exciting testbed for natural language processing, and our 2015 paper is—to the best of our knowledge—the first NLP research on Diplomacy. Before a betrayal, betrayers write ostensibly friendly messages and become more polite, stop talking about the future, and limit how *much* they write [28]. In follow-on work, we developed a dataset that predicts both when users lie to each other and when recipients of lies detect deception [29]. Diplomacy may be a nerdy game, but it is a

fruitful testbed to teach computers to understand messy, emotional human interactions.

Recently, the use of NLP methods in the game of Diplomacy has been the subject of highly-publicized papers by DeepMind in Nature Communications [21] and Meta in Science [2]. The DeepMind paper built a game theoretic understanding of when betrayal should happen, building on our descriptive investigation of deception in human games. The Meta paper, like our 2020 paper, used a classifier to detect deceptive statements but went far beyond our work by building an AI to play Diplomacy. In our most recent work, we showed that even though Meta's *tour de force* agent can consistently beat humans, it is viewed as less trustworthy than human players: humans lie to AI more than other humans and humans view it as lying more (even though it lies less). Moreover, it was less persuasive—i.e., able to change people's minds—than humans despite its strategic supremacy [41].

Persuasion online is not just arguments and interpersonal relations; it's often about *mis*information. Thus, we have focused on developing fact checking: datasets for general knowledge fact checking [8] and climate change fact checking [7]. However, not all misinformation is written so I'm also exploring multimodal misleading information with journalism professor Naeemul Hassan [38].

4. Human-in-the-Loop Adversarial Examples

One of the most fun aspects of my research has been building trivia-playing robots [5, 19, 20]; beyond research papers, our system has faced off against former Jeopardy champions in front of hundreds high school students[2] and against researchers at NeurIPS 2015 (which won the best demonstration award). But after defeating some of the smartest trivia players, did I actually believe that our system was better at question answering? No!

Adversarial examples first came out of the vision community: add a small epsilon to an example and suddenly a object detector calls a turtle a gun [1].[3] While others have attempted to create adversarial examples for language using paraphrasing, it's hard to know if the changes are perceptually negligible ("who wrote the invisible man"—a question with the answer H.G. Wells—is fundamentally different from "who wrote the man you can't see"—an ill-formed questions—as is "who wrote the book invisible man"—a question with the answer Ralph Ellison) and it's hard to "add epsilon" to a discrete word.

Consistent with the theme of my research, my NSF CAREER grant added a *human in the loop* to generate novel adversarial language examples that can provide new training examples to make AI more robust and to expose what AI cannot (yet) do. With Eric Wallace, an undergraduate student, we built a

[2] `https://www.youtube.com/watch?v=LqsUaprYMOw`

[3] Point of personal pride: I mentored Kevin on another research project [15], but I myself had nothing to do with this later adversarial work.

system that could help an expert trivia question writer to stump a computer: as the author writes the question, it shows the author what the system is thinking [39]. And it worked, even generalizing across models [40] (many examples written with an IR model still stump a neural model). After we introduced human-in-the-loop adversarial example generation, Meta/Facebook adopted this framework with gusto [4] in their Dynabench framework, the Dynamic Adversarial Data Collection workshop and call for proposals (which I'm grateful funded our continuing research in this area).

Adversarial examples have become known as "jailbreaks" in the computer security community, and we have also built a dataset with human-in-the-loop adversarial examples that revealed new attach methods such as context filling [30] (this paper won the outstanding Theme Paper award at EMNLP 2023). However, because models are improving so quickly, it is difficult to know how "adversarial" an example is and whether what is adversarial on Monday will remain adversarial on Thursday. Thus, we developed a new metric (AdvScore) based on item response theory to empirically measure how difficult examples are for both humans and humans [37]: adversarial examples are naturally those with the biggest gap between human and computer difficulty.

5. But wait, there's more!

Many of our best-cited papers are "traditional" papers that do better on some task:

- We developed deep averaging networks [20, DAN], a simple model still used in the transformer age [42].
- In question answering, we proposed new evaluation mechanisms for knowing if an answer is correct [33] and improved information retrieval to answer complicated questions [9, 31, 44].
- We also introduced reinforcement learning to *simultaneous machine interpretation* [13], a language-based task that requires significant human intuition, insight, and—for those who want to become interpreters—training.[4] We learned tricks from professional human interpreters—passivizing sentences and guessing the verb—to translate sentences sooner [16], letting speakers and algorithms cooperate together and enabling more natural cross-cultural communication. We also use reinforcement learning to learn machine translation feedback from noisy supervision such as star ratings on a webpage [27].
- We deployed a chatbot to assist new and expectant mothers [36], focusing on techniques to detect questions with incorrect pragmatic assumptions (i.e., a mother asking "when can I resume breastfeeding after a cold" suggests the mother believes they should *stop* breastfeeding during a cold, when that's the only way infants can get the

[4]This framework—using reinforcement learning to capture human strategies—was featured in Liang Huang's ACL keynote.

appropriate antibodies). Many modern language models just accept those false assumptions rather than correct them. We are evaluating the efficacy of this support in a randomized control trial with partners in public health [26].

This work doesn't *yet* fit nicely into the human–computer collaboration narrative, but these more complex tasks are part of my broader vision for where my research will go: state-of-the-art models built to support human decisions, not replace them. And that requires the low-latency models built to react to input "like a human" described above.

6. Future Work

Multimodal. Our adversarial work with language models is nearing an end: it's getting harder and harder to stump them. However, models still struggle with low frequency words when answering spoken questions, understanding abstract representational art, or mapping a schematic to the real world. These leaps of understanding are still a core component of human intelligence, so we will continue to craft adversarial domains but in new modalities to better understand AI limits and highlight human skills that have yet to be mastered.

Backpropagating Cooperation. While we have built metrics that reveal when AIs are effectively cooperating with humans (i.e., raising their skill at a task), most AI models are trained with preference data. My student Nishant Balepur has built a sequence of work showing that what users like isn't always what helps them: for example, students don't always prefer the study aids that best improves retention [3]. The next step is to actually backpropagate that signal into the internals of a model to see what it looks like when the first priority of a model training is to cooperate with users. However, because distinct individuals need different support, this also requires greater personalization.

Data Diversity. Existing datasets are not diverse and do not reflect the kinds of interactions people from diverse backgrounds have with AI systems. In question answering, Google's Natural Questions, SQuAD, and other datasets contain entities that are overwhelmingly male and either American or British [12]. In newly funded NSF research, we're working with surfacing questions that reflect a specific *cultural* context: detecting when questions are answered differently in Ghana than in the US or topics that only people in Bhutan care about (and would be neglected by US-centric datasets).

Proxy Models. While eventually we will need to connect to expensive, difficult to compute neural models for our Human–AI interactions, many user updates can be approximated by fast spectral or probabilistic algorithms. We will design fast, browser-based Javascript approximations of these complicated neural models, to allow users to quickly interact with the models, get a result, and continue before reconciling the solution later.

At a high level, the goal of my research is to build the future I want for my children: an ecosystem not where AI is replacing humans but where it is put in a position to effectively augment their abilities, no matter their skill level or background.

Full list of my publications at
http://boydgraber.org/dyn-pubs/year.html

References

[1] Athalye, A., Engstrom, L., Ilyas, A., Kwok, K.: Synthesizing robust adversarial examples. In: Proceedings of the International Conference of Machine Learning (2018)

[2] Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A.P., Komeili, M., Konath, K., Kwon, M., Lerer, A., Lewis, M., Miller, A.H., Mitts, S., Renduchintala, A., Roller, S., Rowe, D., Shi, W., Spisak, J., Wei, A., Wu, D., Zhang, H., Zijlstra, M.: Human-level play in the game of Diplomacy by combining language models with strategic reasoning. Science 378(6624), 1067–1074 (2022)

[3] Balepur, N., Shu, M., Hoyle, A., Robey, A., Feng, S., Goldfarb-Tarrant, S., Boyd-Graber, J.: A smart mnemonic sounds like "glue tonic": Mixing llms with student feedback to make mnemonic learning stick. In: Proceedings of Empirical Methods in Natural Language Processing (2024), http://cs.umd.edu/~jbg//docs/2024_emnlp_mnemonic.pdf

[4] Bartolo, M., Roberts, A., Welbl, J., Riedel, S., Stenetorp, P.: Beat the AI: Investigating adversarial human annotation for reading comprehension. Transactions of the Association for Computational Linguistics 8, 662–678 (2020)

[5] Boyd-Graber, J., Satinoff, B., He, H., Daume III, H.: Besting the quiz master: Crowdsourcing incremental classification games. In: Proceedings of Empirical Methods in Natural Language Processing (2012)

[6] Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Proceedings of Advances in Neural Information Processing Systems (2009)

[7] Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., Leippold, M.: CLIMATE-FEVER: A dataset for verification of real-world climate claims. In: NeurIPS Workshop on Tackling Climate Change with Machine Learning (2020)

[8] Eisenschlos, J.M., Dhingra, B., Bulian, J., Börschinger, B., Boyd-Graber, J.: Fool me twice: Entailment from wikipedia gamification. In: Conference of the North American Chapter of the Association for Computational Linguistics (2021)

[9] Elgohary, A., Peskov, D., Boyd-Graber, J.: Can you unpack that? learning to rewrite questions-in-context. In: Proceedings of Empirical Methods in Natural Language Processing (2019)

[10] Feng, S., Boyd-Graber, J.: What AI can do for me: Evaluating machine learning interpretations in cooperative play. In: International Conference on Intelligent User Interfaces (2019)

[11] Feng, S., Boyd-Graber, J.: Learning to explain selectively: A case study on question answering. In: Proceedings of Empirical Methods in Natural Language Processing (2022)

[12] Gor, M., Webster, K., Boyd-Graber, J.: Toward deconfounding the influence of subject's demographic characteristics in question answering. In: Proceedings of Empirical Methods in Natural Language Processing (2021)

[13] Grissom II, A., He, H., Boyd-Graber, J., Morgan, J.: Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In: Proceedings of Empirical Methods in Natural Language Processing (2014)

[14] Gu, F., Wongkamjan, W., Kummerfeld, J.K., Peskoff, D., May, J., Boyd-Graber, J.: Personalized help for optimizing low-skilled users' strategy. In: Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (2025), http://cs.umd.edu/~jbg//docs/2024_arr_chiron-advisor.pdf

[15] He, H., Boyd-Graber, J., Kwok, K., Daumé III, H.: Opponent modeling in deep reinforcement learning. In: Proceedings of the International Conference of Machine Learning (2016)
[16] He, H., Grissom II, A., Boyd-Graber, J., Daumé III, H.: Syntax-based rewriting for simultaneous machine translation. In: Proceedings of Empirical Methods in Natural Language Processing (2015)
[17] Hoyle, A., Goel, P., Peskov, D., Hian-Cheong, A., Boyd-Graber, J., Resnik, P.: Is automated topic model evaluation broken?: The incoherence of coherence. In: Neural Information Processing Systems (2021)
[18] Hu, Y., Boyd-Graber, J., Satinoff, B., Smith, A.: Interactive topic modeling. Machine Learning 95(3), 423–469 (Jun 2014), http://dx.doi.org/10.1007/s10994-013-5413-0
[19] Iyyer, M., Boyd-Graber, J., Claudino, L., Socher, R., Daumé III, H.: A neural network for factoid question answering over paragraphs. In: Proceedings of Empirical Methods in Natural Language Processing (2014)
[20] Iyyer, M., Manjunatha, V., Boyd-Graber, J., Daumé III, H.: Deep unordered composition rivals syntactic methods for text classification. In: Proceedings of the Association for Computational Linguistics (2015)
[21] Kramár, J., Eccles, T., Gemp, I., Tacchetti, A., McKee, K.R., Malinowski, M., Graepel, T., Bachrach, Y.: Negotiation and honesty in artificial intelligence methods for the board game of Diplomacy. Nature Communications 13(1) (2022)
[22] Kumar, V., Smith, A., Findlater, L., Seppi, K., Boyd-Graber, J.: Why didn't you listen to me? comparing user control of human-in-the-loop topic models. In: Proceedings of the Association for Computational Linguistics (2019)
[23] Lau, J.H., Newman, D., Baldwin, T.: Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: Proceedings of the European Chapter of the Association for Computational Linguistics (2014)
[24] Li, Z., Mao, A., Stephens, D.K., Goel, P., Walpole, E., Fung, J.F., Dima, A., Boyd-Graber, J.L.: Tenor: Topic enabled neural organization and recommendation: Evaluating topic models in task based settings. In: European Association for Computational Linguistics (2024), http://cs.umd.edu/~jbg//docs/2024_eacl_tenor.pdf
[25] Lund, J., Cook, C., Seppi, K., Boyd-Graber, J.: Tandem anchoring: A multiword anchor approach for interactive topic modeling. In: Proceedings of the Association for Computational Linguistics (2017)
[26] Mane, H.Y., Channell Doig, A., Marin Gutierrez, F.X., Jasczynski, M., Yue, X., Srikanth, N., Mane, S., Sun, A., Moats, R.A., Patel, P., He, X., Boyd-Graber, J., Aparicio, E.M., Nguyen, Q.C.: Practical guidance for the development of rosie, a health education question-and-answer chatbot for new mothers. Journal of Public Health Management and Practice (2023), https://journals.lww.com/jphmp/fulltext/2023/09000/practical_guidance_for_the_development_of_rosie,_a.9.aspx
[27] Nguyen, K., Boyd-Graber, J., Daumé III, H.: Reinforcement learning for bandit neural machine translation with simulated human feedback. In: Proceedings of Empirical Methods in Natural Language Processing (2017)
[28] Niculae, V., Kumar, S., Boyd-Graber, J., Danescu-Niculescu-Mizil, C.: Linguistic harbingers of betrayal: A case study on an online strategy game. In: Proceedings of the Association for Computational Linguistics (2015)
[29] Peskov, D., Cheng, B., Elgohary, A., Barrow, J., Danescu-Niculescu-Mizil, C., Boyd-Graber, J.: It takes two to lie: One to lie and one to listen. In: Proceedings of the Association for Computational Linguistics (2020)

[30] Schulhoff, S.V., Pinto, J., Khan, A., Bouchard, L.F., Si, C., Boyd-Graber, J.L., Anati, S., Tagliabue, V., Kost, A.L., Carnahan, C.R.: Ignore this title and hackaprompt: Exposing systemic vulnerabilities of llms through a global prompt hacking competition. In: Empirical Methods in Natural Language Processing (2023), http://cs.umd.edu/~jbg//docs/2023_emnlp_hackaprompt.pdf
[31] Shi, T., Zhao, C., Boyd-Graber, J., Daumé III, H., Lee, L.: On the potential of lexico-logical alignments for semantic parsing to sql queries. In: Findings of EMNLP (2020)
[32] Shneiderman, B.: Human-Centered AI: A New Synthesis. Springer-Verlag, Berlin, Heidelberg (2021)
[33] Si, C., Zhao, C., Boyd-Graber, J.: What's in a name? answer equivalence for open-domain question answering. In: Proceedings of Empirical Methods in Natural Language Processing (2021)
[34] Smith, A., Boyd-Graber, J., Fan, R., Birchfield, M., Wu, T., Weld, D., Findlater, L.: No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In: International Conference on Human Factors in Computing Systems (2020)
[35] Smith, A., Kumar, V., Boyd-Graber, J., Seppi, K., Findlater, L.: Digging into user control: Perceptions of adherence and instability in transparent models. In: International Conference on Intelligent User Interfaces (2020)
[36] Srikanth, N., Sarkar, R., Mane, H.Y., Aparicio, E.M., Nguyen, Q.C., Rudinger, R., Boyd-Graber, J.: Pregnant questions: The importance of pragmatic awareness in maternal health question answering. In: Conference of the North American Chapter of the Association for Computational Linguistics (2024), http://cs.umd.edu/~jbg//docs/2024_naacl_pregnant.pdf
[37] Sung, Y.Y., Gor, M., Fleisig, E., Mondal, I., Boyd-Graber, J.L.: Advscore: A metric for the evaluation and creation of adversarial benchmarks (2025), http://cs.umd.edu/~jbg//docs/2025_naacl_advscore.pdf
[38] Sung, Y.Y., Hassan, N., Boyd-Graber, J.: Not all fake news is written: A dataset and analysis of misleading video headlines. In: Proceedings of Empirical Methods in Natural Language Processing (2023), http://cs.umd.edu/~jbg//docs/2023_emnlp_videoheadline.pdf
[39] Wallace, E., Boyd-Graber, J.: Trick me if you can: Adversarial writing of trivia challenge questions. In: ACL Student Research Workshop (2018)
[40] Wallace, E., Rodriguez, P., Feng, S., Yamada, I., Boyd-Graber, J.: Trick me if you can: Human-in-the-loop generation of adversarial question answering examples. Transactions of the Association for Computational Linguistics 10 (2019)
[41] Wongkamjan, W., Gu, F., Wang, Y., Hermjakob, U., May, J., Stewart, B.M., Kummerfeld, J.K., Peskoff, D., Boyd-Graber, J.L.: More victories, less cooperation: Assessing cicero's diplomacy play. In: Proceedings of the Association for Computational Linguistics (2024), http://cs.umd.edu/~jbg//docs/2024_acl_cicero.pdf
[42] Ye, Q., Khabsa, M., Lewis, M., Wang, S., Ren, X., Jaech, A.: Sparse distillation: Speeding up text classification by using bigger student models. In: Conference of the North American Chapter of the Association for Computational Linguistics (2022)
[43] Yuan, M., Van Durme, B., Boyd-Graber, J.: Multilingual anchoring: Interactive topic modeling and alignment across languages. In: Neural Information Processing Systems (2018)
[44] Zhao, C., Xiong, C., Boyd-Graber, J., Daumé III, H.: Distantly-supervised dense retrieval enables open-domain question answering without evidence annotation. In: Proceedings of Empirical Methods in Natural Language Processing (2021), http://umiacs.umd.edu/~jbg//docs/2021_emnlp_weak_dpr.pdf