Detecting and Addressing *BIAS* in Data, Humans, and Institutions

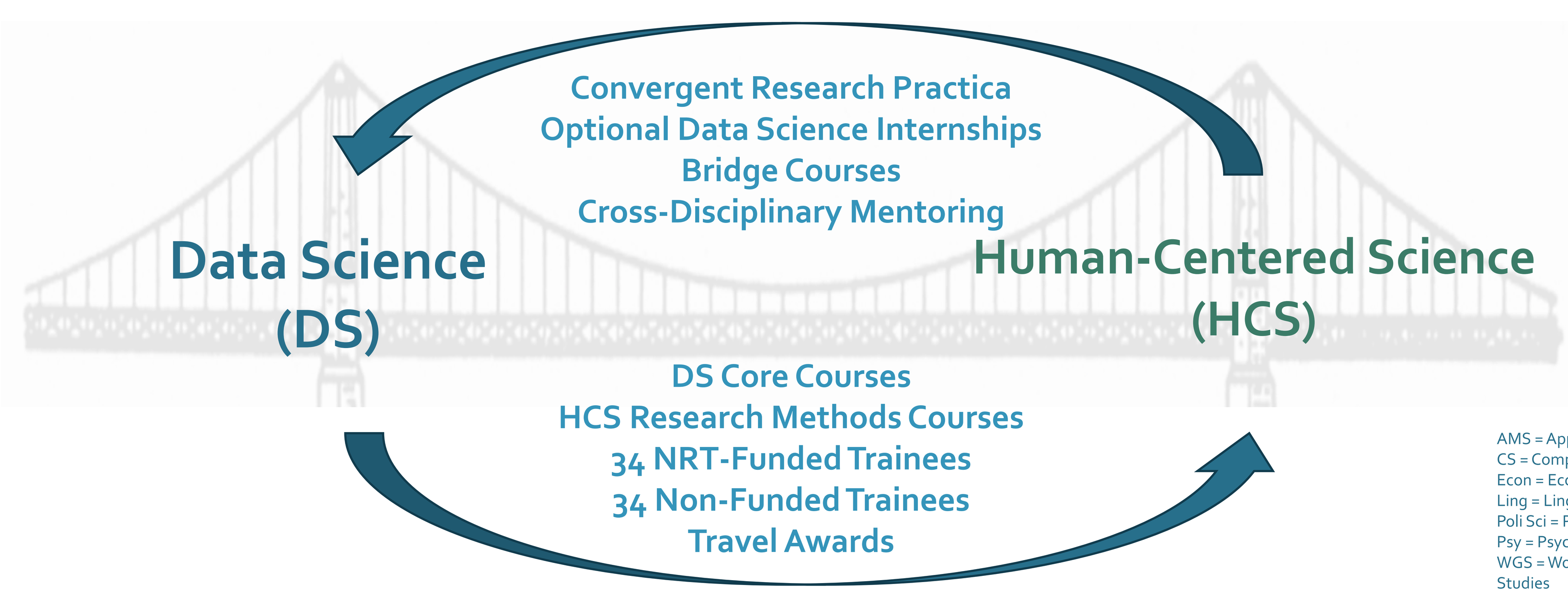
About

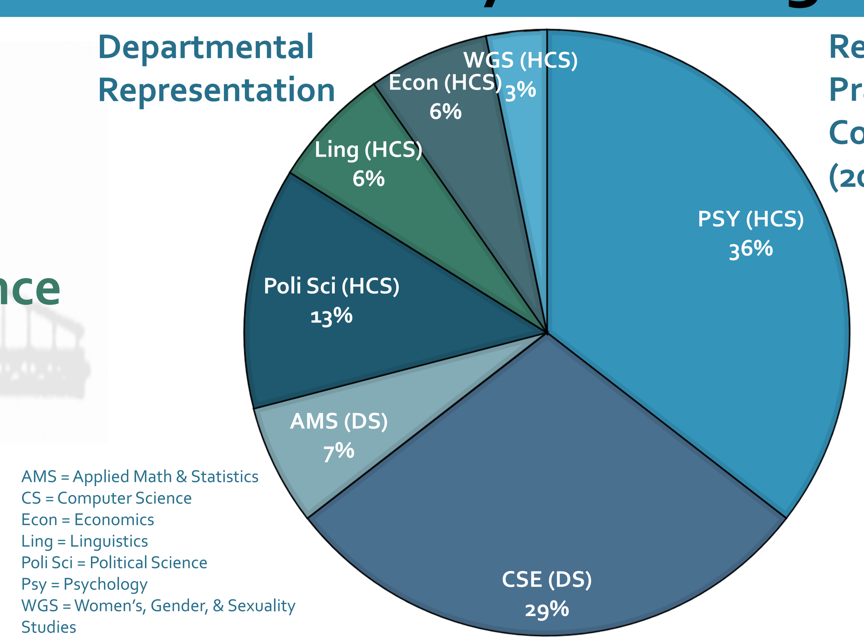
Data science and AI are powerful tools for generating new knowledge, fueling innovation, and dealing with society's most pressing problems. However, "big data" and machine learning tools can perpetuate biases that advantage some people, and disadvantage others. This training project (NSF 2125295) bridges perspectives from the human-centered sciences with those from the data sciences in support of convergent research projects.

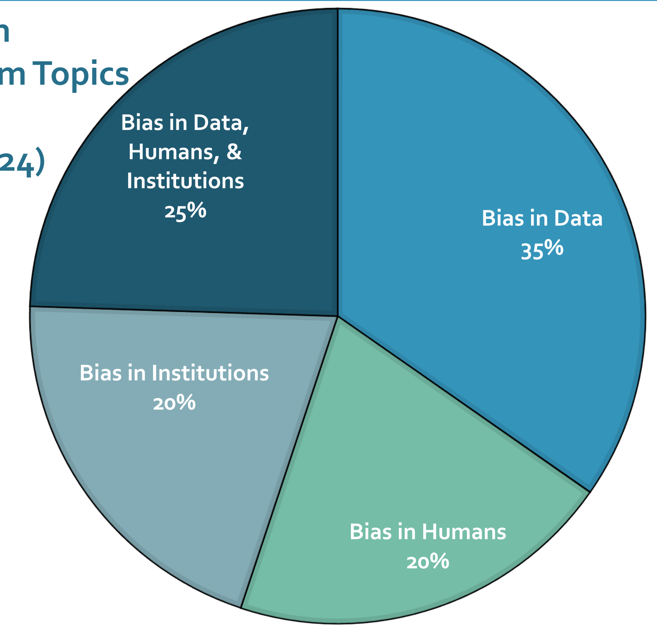
Back row: Susan E. Brennan (PI), Jeffrey Heinz (Co-PI), Adryan Wallace (Bias-NRT Faculty), CR Ramakrishnan (Co-PI), and Reuben Kline (Bias-NRT Faculty)
Front row: Wei Zhu (Co-PI), , Bonita London (Co-PI), , and Owen Rambow (Bias-NRT Faculty)

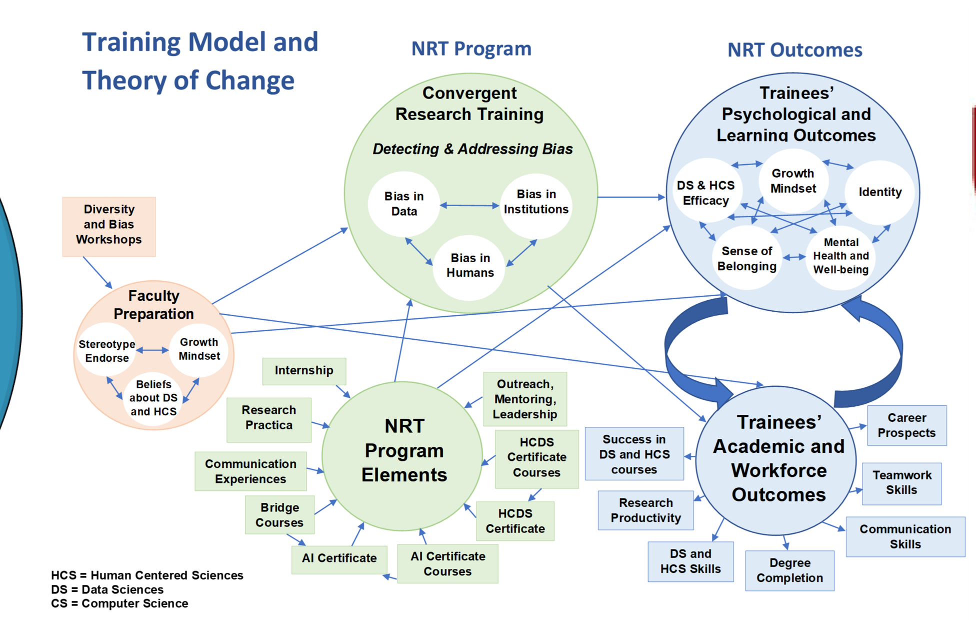
Leadership Team: Susan E. Brennan (PI), C.R. Ramakrishnan, Wei Zhu, Bonita London, Jeffrey Heinz
Project Coordinator: Kristen Kalb-DellaRatta
Project Evaluation: Catherine Good, Elevate Learning, LLC

Mission: to seed a generation of researchers trained to identify and mitigate biases that arise when data-centric methods are applied to real-world problems









This project is based on work supported by NSF Grant #2125295

Trainee Research Highlights

Training LLMs to Recognize Hedges in Spontaneous Narratives

Amie J. Paige, Adil Soubki, John Murzaku, Owen Rambow, & Susan E. Brennan

Bias-NRT Trainees, Amie Paige, John Murzaku, and Adil Soubki presented a poster for their paper, ***Training LLMs to Recognize Hedges in Spontaneous Narratives*** at SIGDIAL 2024 in Kyoto, Japan

Abstract

Hedges allow speakers to mark utterances as provisional, whether to signal non- prototypicality or "fuzziness", to indicate a lack of commitment to an utterance, to attribute responsibility for a statement to someone else, to invite input from a partner, or to soften critical feedback in the service of face-management needs. Unlike humans, current LLMs use hedges indiscriminately. Here we focus on hedges in an experimentally parameterized from naturalistic storytelling dialogues (Galati and Brennan, 2010). First, we coded the corpus for hedges. Then, we compared commercial LLMs hedge detection performance against a smaller fine-tuned BERT model with various prompting strategies.

Model	Training	Prompt	Precision (P)	Recall (R)	F1 Score (F1)
BERT	Finetuned	-	0.883 ± 0.015	0.934 ± 0.012	0.908 ± 0.010
GPT-4o	Few-Shot	List	0.613 ± 0.027	0.848 ± 0.018	0.712 ± 0.021
LLaMA-3	Few-Shot	List	0.518 ± 0.035	0.799 ± 0.022	0.628 ± 0.031
GPT-4o	Few-Shot	BIO	0.514 ± 0.024	0.766 ± 0.036	0.616 ± 0.030
GPT-4o	Zero-Shot	List	0.430 ± 0.014	0.711 ± 0.004	0.536 ± 0.012
GPT-4o	Zero-Shot	BIO	0.436 ± 0.026	0.618 ± 0.033	0.510 ± 0.028
LLaMA-3	Few-Shot	BIO	0.298 ± 0.018	0.625 ± 0.016	0.404 ± 0.019
LLaMA-3	Zero-Shot	BIO	0.167 ± 0.014	0.428 ± 0.019	0.240 ± 0.017
LLaMA-3	Zero-Shot	List	0.274 ± 0.023	0.146 ± 0.010	0.190 ± 0.011

Table 2: Average performance metrics over the five folds with standard deviations for different models, training methods, and prompt types, ordered by F1 score.

The BERT model greatly outperformed the commercial models, suggesting that additional experience with naturalistic dialogues that contain hedges could improve commercial models. Our work paves the way for both multimodal hedge detection tasks and hedge generation tasks using LLMs.

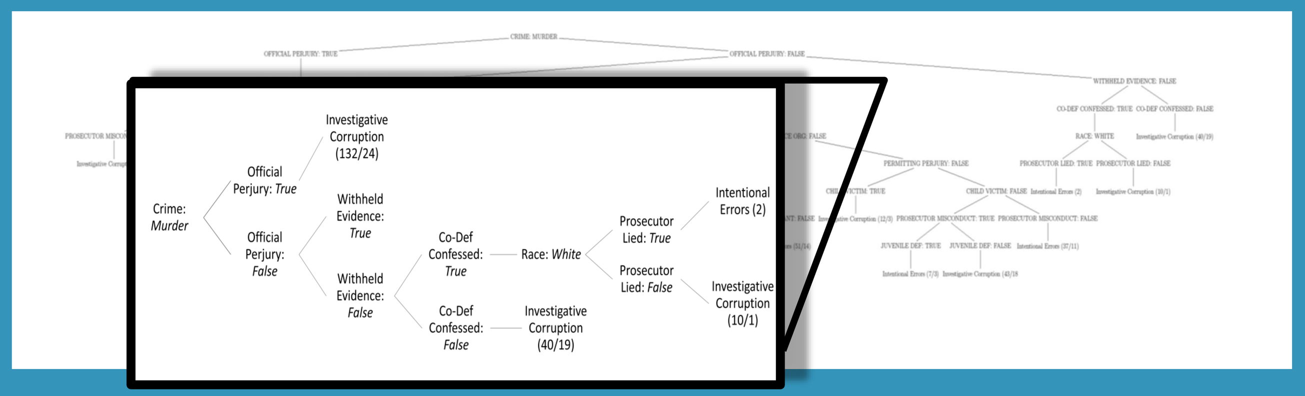
Paige, A. J., Soubki, A., Murzaku, J., Rambow, O., & Brennan, S. E. (2024). Training llms to recognize hedges in dialogues about Roadrunner cartoons. *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 204–215. https://doi.org/10.18653/v1/2024.sigdial-1.18

A Computational Decision-Tree Approach to Inform Post-Conviction Intake Decisions

Kalina Kostyszyn, Carl J. Wiedemann, Rosa Bermejo, Amie Paige, Kristen W. Kalb-DellaRatta, & Susan E. Brennan

Abstract

How might data analytic tools support intake decisions? When faced with a request for post-conviction assistance, innocence organizations' intake staff must determine (1) whether the applicant can be shown to be factually innocent, and (2) whether the organization has the resources to help. These difficult categorization decisions are often made with incomplete information (Weintraub, 2022). We explore data from the National Registry of Exonerations (NRE; 4/26/2023, N = 3,284 exonerations) to inform such decisions, using patterns of features associated with successful prior cases. We first reproduce Berube et al. (2023)'s latent class analysis, identifying four underlying categories across cases. We then apply a second technique to increase transparency, decision tree analysis (WEKA, Frank et al., 2013). Decision trees can decompose complex patterns of data into ordered flows of variables, with the potential to guide intermediate steps that could be tailored to the particular organization's limitations, areas of expertise, and resources.



Above, a Decision Tree trained on exoneration data to predict trends associated with latent class membership. This branch organizes cases marked as 'murders.' Depending on the features associated with each case, the case is labeled as one of the latent classes.

Bias-NRT Trainees and lead authors, Kalina Kostyszyn and Carl Wiedemann, presented their paper, *A Computational Decision-Tree Approach to Inform Post-Conviction Intake Decisions*, at the Just Data 2023: Advancing the Innocence Movement conference on November 9th, 2023, and were published in the *Wrongful Conviction Law Review*

INNOCENCE PROJECT

Kostyszyn, K., Wiedemann, C., Bermejo, R., Paige, A., Kalb-DellaRatta, K., & Brennan, S. (2024). A computational decision-tree approach to inform post-conviction intake decisions. *The Wrongful Conviction Law Review*, 5(1), 80–102. https://doi.org/10.29173/wclawr110

Fall 2024 Welcome Event

Back Row: Zhengxiang Wang (Linguistics), Reuben Kline (Bias-NRT Faculty, Political Science), Medhini Urs (Psychology, Cognitive Science), C.R. Ramakrishnan (Co-PI, Computer Science), Owen Rambow (Bias-NRT Faculty, Linguistics), Brett Indelicato (Economics), Evan West (Computer Science), and Gilvir Gill (Computer Science). **Middle Row**: Amie Paige (Psychology, Cognitive Science), Susan Brennan (PI, Psychology, Cognitive Science), Amit Kumar Das (Computer Science), Adil Soubki (Computer Science), Karin Hasegawa (Applied Math & Statistics), Dasha Likhacheva (Psychology, Social & Health), Ritik Raina (Psychology, Cognitive Science), Ignacio Urbina (Political Science), Rosa Bermejo (Psychology, Social & Health), Carl Wiedemann (Psychology, Social & Health), Sri Jangili (Political Science), Tina Behzad (Computer Science), Alexandra Anthonioz (Psychology, Social & Health), Kiera Gross (Computer Science), Benjy Hechtman (Applied Math & Statistics), MacKenzie Johnson (Psychology, Cognitive Science), and James May (Psychology, Cognitive Science). **Front Row**: Kristen Kalb-DellaRatta (Project Coordinator)

Advanced Graduate Certificate in Human-Centered Data Science

Recently approved by the State University of New York (SUNY) and the New York State Education Department (NYSED), the **Advanced Graduate Certificate in Human-Centered Data Science** (HCDS) is now available for enrollment. Trainees, Fellows, and other PhD students from eight participating departments are eligible to enroll. The certificate requires 12-credits (four courses): two core data science/computer science courses and two human-centered science electives. In addition to the 12-credits, all students enrolled in the HCDS certificate will complete the online Citi IRB Training, "Human Research." As of the Spring 2024 semester, **48% of trainees were enrolled** in the certificate track and by the semester's end, **four trainees had completed it.**

State University of New York

State University at Stony Brook

On the recommendation of the Faculty and by virtue of the authority vested in them the Trustees of the University have conferred on

Tina Behzad

the Advanced Graduate Certificate in

Human-Centered Data Science

Given at Stony Brook, in the State of New York, in the United States of America on the seventeenth day of May two thousand twenty-four.

Learn More!

Poster created by Kristen Kalb-DellaRatta, Project Coordinator