

# Peter Hase

peter@cs.unc.edu · [peterbhase.github.io](https://peterbhase.github.io) · (919) 323-0393

## EDUCATION

### The University of North Carolina at Chapel Hill

Third-year PhD student in Computer Science

Research Area: Natural Language Processing | Advisor: [Mohit Bansal](#)

Fall 2019 – Present

Chapel Hill, NC

### Duke University

BS in Statistical Science | Minor in Mathematics

Fall 2015 – Spring 2019

Durham, NC

## RESEARCH INTERESTS

Interpretable and explainable machine learning, natural language processing,  
multi-agent communication, AI safety.

## PUBLICATIONS

### Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov,  
Mohit Bansal, Srinivasan Iyer

Preprint on arXiv. [[pdf](#)] [[code](#)]

### Low-Cost Algorithmic Recourse for Users With Uncertain Cost Functions

Prateek Yadav, Peter Hase, Mohit Bansal

Preprint on arXiv. [[pdf](#)] [[code](#)]

### Search Methods for Sufficient, Socially-Aligned Feature Importance Explanations with In-Distribution Counterfactuals

Peter Hase, Harry Xie, Mohit Bansal

In *NeurIPS 2021*. [[pdf](#)] [[code](#)]

### When Can Models Learn From Explanations? A Formal Framework for Understanding the Roles of Explanation Data

Peter Hase, Mohit Bansal

Preprint on arXiv. [[pdf](#)] [[code](#)]

### FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging

Han Guo, Nazneen Fatema Rajani, Peter Hase, Mohit Bansal, Caiming Xiong

In *EMNLP 2021*. [[pdf](#)] [[code](#)]

### Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?

Peter Hase, Shiyue Zhang, Harry Xie, Mohit Bansal

In *Findings of EMNLP 2020*. [[pdf](#)] [[code](#)]

### Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?

Peter Hase, Mohit Bansal

In *ACL 2020*. [[pdf](#)] [[code](#)]

### Interpretable Image Recognition with Hierarchical Prototypes

Peter Hase, Chaofan Chen, Oscar Li, Cynthia Rudin

In *AAAI-HCOMP 2019*. [[pdf](#)] [[code](#)]

## Shall I Compare Thee to a Machine-Written Sonnet? An Approach to Algorithmic Sonnet Generation

John Benhardt, Peter Hase, Liuyi Zhu, Cynthia Rudin

Preprint on arXiv. [[pdf](#)] [[code](#)]

### AWARDS

**Google PhD Fellowship (Natural Language Processing)**, Google 2021

Fellowship awarded to six students globally for research in Natural Language Processing, providing up to three years of full funding

**Royster PhD Fellowship**, UNC Chapel Hill 2019

University fellowship awarded to one student in the 2019 cohort of computer science students, providing three years of full funding

**First Prize in the PoetiX Literary Turing Test**, Neukom Institute, Dartmouth College 2018

Awarded for the top submission to the Neukom Institute's open competition for algorithmic sonnet generation

**Nomination for Undergrad TA of the Year**, Dept. of Statistical Science, Duke University 2018

One of five undergrad nominations from faculty for the department's TA of the year award

**ASA DataFest Honorable Mention**, Dept. of Statistical Science, Duke University 2018

Recognition for placement in top 10% of teams in a Duke-hosted data analysis competition entered by 380+ undergrad and grad students

**Meritorious Winner in the Interdisciplinary Contest in Modeling**, COMAP 2017

Awarded for placement in the top 12% of over 8000 teams in the international modeling contest held by the Consortium for Mathematics and its Applications

**A.J. Tannenbaum Trinity Scholarship**, Duke University 2015

A full academic merit scholarship awarded to one student from Guilford County, NC

### INVITED TALKS

**NEC Laboratories Europe** Spring 2022

"Explainable Machine Learning in NLP: Methods and Evaluation" [[slides](#)]

**National Institute for Standards and Technology (NIST)** Spring 2022

"Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?" [[slides](#)]

**Allen Institute for AI** Spring 2022

"Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs?" [[slides](#)]

**Uber AI** Spring 2022

"The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations" [[slides](#)]

**Center for Human Compatible AI (CHAI), UC Berkeley** Summer 2021

"Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?" [[slides](#)]

### TEACHING

**Probabilistic Machine Learning (Graduate)**, Teaching Assistant Spring 2019

Dept. of Statistical Science, Duke University

<b>Intro to AI</b> , Teaching Assistant Dept. of Computer Science, Duke University	<i>Spring 2019</i>
<b>Elements of Machine Learning</b> , Teaching Assistant Dept. of Computer Science, Duke University	<i>Fall 2018</i>
<b>Intro to Data Science</b> , Teaching Assistant Dept. of Statistical Science, Duke University	<i>Spring 2018</i>
<b>Regression Analysis</b> , Teaching Assistant Dept. of Statistical Science, Duke University	<i>Fall 2017</i>

**PROFESSIONAL  
SERVICE**

<b>Program Committees</b> Reviewer	<i>Summer 2020 – Present</i>
<ul style="list-style-type: none"> <li>• ACL Rolling Review, February 2022</li> <li>• ACL Rolling Review, January 2022</li> <li>• ACL Rolling Review, December 2021</li> <li>• ACL Rolling Review, October 2021</li> <li>• ACL Rolling Review, September 2021</li> <li>• NeurIPS DistShift Workshop 2021</li> <li>• EMNLP BlackboxNLP Workshop 2021</li> <li>• EMNLP 2021</li> <li>• ACL-IJCNLP 2021 (<i>Outstanding Reviewer</i>)</li> <li>• ICLR RobustML Workshop 2021</li> <li>• NAACL-HLT 2021</li> <li>• EACL 2021</li> <li>• EMNLP 2020 (<i>Outstanding Reviewer</i>)</li> </ul>	

**RESEARCH  
EXPERIENCE**

<b>Meta AI Research</b> Research Intern   <i>Supervisor:</i> Dr. Srinivasan Iyer	<i>Summer 2021 Seattle, WA</i>
<ul style="list-style-type: none"> <li>• Worked on methods for detecting and updating beliefs/knowledge in language models</li> <li>• Produced paper on the topic, “Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs”</li> </ul>	
<b>Department of Statistical Science, Duke University</b> DOmath Researcher   <i>Supervisor:</i> Dr. Sayan Mukherjee	<i>Summer 2018 Durham, NC</i>
<ul style="list-style-type: none"> <li>• Numerically estimated a measure of model complexity, the topological entropy, for two dynamical systems, the logistic map and linear dynamical system</li> <li>• Empirically assessed how the reliability of inference for the linear dynamical system varies as a function of its entropy</li> </ul>	
<b>Department of Neurobiology, Duke University</b> Research Assistant   <i>Supervisor:</i> Dr. Jeff Beck	<i>Spring &amp; Summer 2018 Durham, NC</i>
<ul style="list-style-type: none"> <li>• Implemented a hidden Markov model and linear dynamical system, each learned through variational Bayesian expectation maximization (VBEM)</li> <li>• Modeled recordings of neuron activity in the actively singing Zebra finch; visualized and interpreted models’ latent variable dynamics</li> </ul>	
<b>Information Initiative at Duke</b> Data+ Researcher   <i>Supervisor:</i> Dr. Sheng Jiang	<i>Summer 2017 Durham, NC</i>

- Clustered Duke's alumni donors into groups with distinct giving behaviors via k-means
- Built logistic regression models to evaluate donors' philanthropic potential based on demographics and prior giving behavior

## LEADERSHIP

### **Computer Science Student Association**

*Summer 2020 – Present*

Officer

*Chapel Hill, NC*

- Observed faculty teaching to provide feedback in tenure review
- Record meeting minutes for CS faculty meetings to share with graduate students
- Working with faculty to streamline background requirements for doctoral applicants

### **High school and Undergraduate Research Mentoring**

*Spring 2020 – Present*

Research Mentor

*Chapel Hill, NC*

- Meet weekly with an undergraduate research assistant in the MURGe-Lab to mentor ongoing publication track research
- Met weekly with a high school student from North Carolina School of Science and Math to mentor a summer project reimplementing current research in document summarization
- Presented live research demos to Chapel Hill K-12 students for UNC CS open house; printed machine written sonnets for students and discussed education and research at UNC

### **Start-up Technical Advising**

*Fall 2019 – Present*

Technical Advisor

*Chapel Hill, NC*

- [curalens.ai](#): previously advised Curalens on text generation strategies for a therapeutic chat-bot (note: Curalens also advised by domain experts)
- [Acta](#): previously advised Acta on procedures for automatically summarizing crowdsourced constituent feedback for efficient communication to local governments

### **Effective Altruism: Duke**

*Spring 2016 – Spring 2019*

Co-President

*Durham, NC*

- Moderated weekly discussions related to Effective Altruism, the social movement promoting the use of reason and evidence to maximize the good you can do for the world
- Organized lectures and reading groups on AI safety for Duke and UNC Chapel Hill students
- Led club from 9 to 30+ active members over my tenure as Co-President
- Recorded over 15 Giving What We Can pledges (10% of all future income) in pledge drives and over 30 One For the World pledges (1% of future income)

## WORK EXPERIENCE

### **Clarity Campaign Labs**

*Summer 2016*

Research Analyst

*Washington, DC*

- Visualized model predictions and political data; encoded surveys; drafted software guides for internal use