Human-Centered Evaluation of Explanations

Jordan Boyd-Graber[1], **Samuel Carton**[2,3], **Shi Feng**[3], **Q. Vera Liao**[4], **Tania Lombrozo**[5], **Alison Smith-Renner**[6], **Chenhao Tan**[3]
[1]University of Maryland, [2]University of New Hampshire, [3]University of Chicago, [4]Microsoft Research, [5]Princeton University, [6]Dataminr

Abstract

The NLP community are increasingly interested in providing explanations for NLP models to help people make sense of model behavior and potentially improve human interaction with models. In addition to computational challenges in generating these explanations, evaluations of the generated explanations require human-centered perspectives and approaches. This tutorial will provide an overview of human-centered evaluations of explanations. First, we will give a brief introduction to the psychological foundation of explanations as well as types of NLP model explanations and their corresponding presentation, to provide the necessary background. We will then present a taxonomy of human-centered evaluation of explanations and dive into depth in the two categories: 1) evaluation with human-subjects studies; 2) evaluation based on human-annotated explanations. We will conclude by discussing future directions. We will also adopt *a flipped format* to maximize the interactive components for the live audience.

Type of Tutorial: It will be designed to provide introductory content for computer scientists, but aim to cultivate cutting-edge interdisciplinary research to work on this inherently human-centric topic by introducing perspectives and methods from psychology and human-computer interaction (HCI).

1 Tutorial Description

Thanks to recent advances in deep learning and large-scale pretrained language models, NLP systems have demonstrated impressive performance in a wide variety of tasks, ranging from classification to generation. However, in order to effectively use these NLP systems in support of human endeavour, it is critical that we can explain model predictions in ways that humans can easily comprehend. Such explanations are particularly important for high-stakes decisions such as hiring and loan approval. Indeed, the NLP community have developed a battery of algorithms and models for explaining model predictions and there have been past tutorials dedicated to such algorithms (Wallace et al., 2020).

However, there is less consensus on how to evaluate explanations. And, since these explanations eventually serve the needs of humans, it is important to take a human-centered approach to their evaluation, meaning evaluating with respect to human criteria, measuring human perceptions of explanations and whether explanations serve human needs. Therefore, interdiscplinary perspectives are necessary for the success of such evaluations, especially ones from psychology and HCI, which is unfamiliar to the NLP community. This tutorial aims to fill in this gap and introduce the nascent area of human-centered evaluation of explanations.

This tutorial will first present the psychological and philosophical foundations of explanations. We will highlight that explanations are heterogeneous and selective. We will discuss diverse goals people seek explanations for, highlighting that effective explanations identify a difference maker, which is often causal. These discussions will lay the foundation for the rest of the tutorial.

We will then introduce the basic elements of explanations and their presentation, including explanation types and taxonomies, so that participants are familiar with the subject of evaluation. We will proceed with a taxonomy of human-centered evaluations, to include two primary types: application-grounded human-subjects evaluations and evaluations based on human-provided explanations.

We start with how to conduct *human-subjects studies* to evaluate explanations. We would like to encourage NLP researchers to move beyond using simplified evaluation tasks, to considering different usage scenarios of explantions and articulating evaluation goals—for whom and what purposes a given explanation method is meant to serve, then define the evaluation task, evaluation criteria, and

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics
Human Language Technologies: Tutorial Abstracts, pages 26 - 32

recruiting requirement accordingly. We will also describe common methods to measure different evaluation criteria, such as survey scales and behavioral measurement, while raising limitations of existing methods.

Then, we cover evaluations with *evaluation based on human-annotated explanations*, as this area is more familiar with the NLP audience. This family of evaluations involves collecting explanations alongside ground-truth labels, and using these human-annotated explanations as a gold standard for model-generated explanations. While intuitive, this practice has validity issues associated with misalignment between human reasoning and model behavior, which we will discuss at length.

We will conclude the tutorial with a discussion of future directions for human-centered evaluation of explanations.

Flipped format. Our tutorial will be in a flipped format: participants view the videos asynchronously and participate in Q&A and work through hands-on activities. The flipped classroom has shown better retention than traditional instruction in a stand-alone instruction session (Bishop and Verleger, 2013). We believe the flipped format is also condusive for ACL tutorials: (1) it will have more longevity, as the recorded (and edited) videos will be of higher quality than videos recorded at a typical conference session; (2) it will be easier for hybrid participation; (3) it will be a more engaging experience for in-person participants.

All of the videos will be in segments of twenty minutes or less for easy asynchronous viewing. To ensure accessibility, we will have manual (not ASR) captions and distribute the slide source along with the videos for easier incorporation of the tutorial into classroom instruction.

Target Audience and Expected Pre-requisite. We welcome anyone who is interested in intepretable NLP and human-AI interaction and only require basic knowledge to programming and contemporary classification models.

2 Outline of the Tutorial Content and Reading List

The tutorial will consiste of two parts: (1) (offline) two hours of content to be viewed asynchronously and (2) (online or in-person) three hours of Q&A and hands-on activities. We include the cited references in the outline description.

2.1 Asynchronous Tutorial

Introduction. This section will introduce explainable AI (XAI) and the importance of evaluating explanations following a human-centered approach (i.e., evaluating with respect to stakeholder needs and desiredata).

Psychological foundation of explanations. This section will cover the research on human explanations in psychology that highlights the fact that human explanations are necessarily incomplete: we do not start from a set of axioms and present all the deductive steps. We will also explore the assumption on whether humans can provide explanations. Furthermore, to build the foundation for defining evaluation goals and criteria for model explanations, we will discuss the diverse goals people seek explanation for. Cited references: Aronowitz and Lombrozo (2020); Aslanov et al. (2021); Blanchard et al. (2018); Giffin et al. (2017); Wilson and Keil (1998); Hemmatian and Sloman (2018); Keil (2003); Kuhn (2001); Lipton (1990); Lombrozo (2012, 2016); Lombrozo et al. (2019); Woodward and Ross (2021).

Explanation methods. The design of evaluation studies is a primary focus of this tutorial. And the subject of these user studies is machine explanations. This section provides the necessary background knowledge on the generation and presentation of machine explanations. We will present a high-level taxonomy of explanation methods and the challenges each category presents to the evaluation. We cover both local explanations such as feature attribution (Ribeiro et al., 2016; Lundberg and Lee, 2017; Li et al., 2016) and counterfactuals (Goyal et al., 2019; Verma et al., 2020), and global explanations such as prototypes (Snell et al., 2017; Gurumoorthy et al., 2019) and adversarial rules (Ribeiro et al., 2018; Wallace et al., 2019). Our overview will omit technical details such as how to computate the input gradient for a specific neural network architecture. Instead, we will discuss the various design choices behind the presentation of explanations, such as color mapping, interactivity, and customizability. For example, local feature importance might be presented as highlighted words in a text classifier, whereas model uncertainty (or prediction probability) can be exposed as either a numerical value or pie chart. Explanations may be provided either alongside every

prediction or only on demand. Explanations might be static information displays or interactive, supporting drilling in for more detail, questioning the system, or even providing feedback to improve it. We will also discuss the limitation of these explanation methods (Guo et al., 2017; Feng et al., 2018; Ye et al., 2021).

Evaluating explanations . We will then provide an overview of human-centered evaluation approaches.

AppliHuman-subjects evaluation . We will start by distinguishing between application-grounded evaluation, based on the success of target users' end goal, and simplified evaluation, such as asking people to simulate the model predictions based on its explanations (Doshi-Velez and Kim, 2017). While it is currently more common for NLP researchers to use simplified evaluation tasks, a recent HCI study pointed out their limitations and lack of evaluative power to predict the actual success in deployment (Buçinca et al., 2020). To encourage NLP researchers to move towards performing application-grounded evaluation, and in a principled and efficient fashion, we will introduce a taxonomy of common applications of explanations, user types and user goals (e.g., model diagnosis, decision improvement, trust calibration, auditing for biases) based on recent HCI work (Suresh et al., 2021; Liao et al., 2020). Using this framework, NLP researchers can articulate the user type(s) and user goal(s) that a given explanation method is meant to serve, and based on that define the evaluation tasks, criteria, subjects to recruit, and so on. We will cover common evaluation criteria regarding both the reception of explanations (e.g., easiness to understand, cognitive workload) and satisfaction of users' end goals, and discuss existing methods to measure them, such as survey scales and behavioral measures. We will also provide introductory contents on how to conduct human-subjects studies, such as how to recruit participants, design tasks and instructions, prevent data noises and biases, and common ethical concerns. We will also give case studies such as Dodge et al. (2019) and Lai and Tan (2019). This tutorial aims to promote important considerations in this nascent area and introduce existing methods from HCI to inspire establishing best practices. Additional references: (Liao and Varshney, 2021; Zhang et al., 2020; Wang and Yin, 2021; McKnight et al., 2002; Cheng et al., 2019; Lai et al., 2021; Kaur et al., 2020; Jacobs et al., 2020).

Evaluation based on human-provided explanations. We discuss the advantages and disadvantages of human-annotated explanations as a means for evaluating model explanations.

Numerous NLP datasets have been released with both labels and human-provided explanations. These come mostly in the form of *rationales* indicating which tokens within a text are important or causal for the true label, e.g., (Zaidan et al., 2007; Khashabi et al., 2018; Thorne et al., 2018), but sometimes consist of *natural language* e.g., (Camburu et al., 2018). DeYoung et al. (2019) aggregates several such datasets into one collection, while Wiegreffe and Marasović (2021) gives an overview of these datasets in the wider literature.

We discuss the metrics by which human-annotated explanations are used to evaluate model-generated explanations. This is a relatively straightforward sequence classification-style evaluation for rationale-type explanations (F1, MSE, etc.), but a more nuanced NLG-style evaluation for natural language explanations (Garbacea and Mei, 2020).

We conclude with a discussion of the validity of human-explanations as a gold standard for model explanations. Recent work has investigated the informational properties of human-annotated explanations, finding that there are gaps between what information humans believe is sufficient or necessary for prediction (i.e. human-annotated explanations), and what actually is so in practice for trained NLP models Carton et al. (2020); Hase et al. (2020). We discuss the implications of these analyses on the validity question, as well as on the future of this style of evaluation.

Summary and future directions . We will conclude by comparing these two main types of human-centered evaluations, recommending best practices, and discussing future directions.

2.2 Q&A and Tutorial Activities

For the in-person tutorial, we will provide a brief recap of the tutorial, followed by an interactive Q&A session and working group activities. We will choose two tasks based on pre-conference surveys as running examples, e.g., sentiment analysis and question answering. Please see the outline below.

- Recap (40 minutes).
- Q&A (40 minutes).

- Break (10 minutes).
- Activity 1: Get familiar with explanations (30 minutes). The main of this exercise is to get them to see how different explanation methods work in practice. We will provide a notebook and models to be used.
- Activity 2: Hands-on participatory evaluation (60 minutes). We will have two tracks that are aligned with the two approaches, one for explanation dataset collection, one for human subject evaluation. It has two steps: 1) research design and 2) participatory evaluation. In this first step, we will ask people to either come up with annotation guideline or articulate evaluation goals (e.g., what user goal(s) and user type(s) is it meant to serve) and define evaluation criteria (e.g., evaluation tasks and measurements). In the second step, parcipants will exchange and participate in the study designed in the first step by either annotating explanations based on the guidelines or performing user studies based on the tasks.

3 Expected Outcome

We plan to make tutorial presentation materials public. We will make sure the videos are accessible to a wide population, e.g., via transcripts.

Estimated audience size. We estimate that ∼200 people will attend the tutorial. The algorithmic counterpart, Wallace et al. (2020), was one of the most popular tutorials at EMNLP that year.

4 Diversity Considerations

Our speakers are diverse in discipline (NLP, HCI, and psychology), gender (4 male, 3 female), seniority (from professor to postdocs), academia and industry (5 from acadmia, 2 from industry).

Our flipped format will accomodate a diverse group of audience because of its asychronous nature. For example, non-native speakers have more time to digest the content. We also require a low barrier of entry. To further attract a diverse group of participants, we will advertise this through underrepresented groups such as Women in NLP, Black in AI, and Queer in AI.

5 Presenter Biographies

Jordan Boyd-Graber is an associate professor at the University of Maryland, with joint appointments between computer science, the iSchool, language science, and the Institute for Advanced Computer Studies. He has been teaching using a flipped classroom approach since 2013. He and his collaborators helped end the use of perplexity for topic models (Chang et al., 2009), first developed interactive topic models (Hu et al., 2011), and improved word-level analysis of topic model explanations (Lund et al., 2019). Additional information at: http://boydgraber.org.

Samuel Carton is a postdoctoral researcher at the University of Colorado, Boulder. His interests lie in model interpretability and human-AI interaction. Additional information at: https://shcarton.github.io.

Shi Feng is a postdoctoral researcher at the Uhiversity of Chicago. His research interests include interpretable NLP, adversarial robustness, and alignment. Additional information at: http://www.shifeng.umiacs.io/.

Q. Vera Liao is a Principal Researcher at Microsoft Research Montreal, where she is part of the FATE (Fairness, Accountability, Transparency, and Ethics) group. She is an HCI researcher by training, with current interest in human-AI interaction and explainable AI. More information can be found at: http://www.qveraliao.com/

Tania Lombrozo is the Arthur W. Marks '19 Professor of Psychology at Princeton University. She is a leading expert in understanding explanations. Additional information is available at: http://cognition.princeton.edu/.

Alison Smith-Renner is a Senior Research Scientist in human-AI interaction at Dataminr. Her research interests include explainable and interactive natural language processing from a human-centric perspective. Additional information is available at: https://alisonmsmith.github.io

Chenhao Tan is an assistant professor of computer science at the University of Chicago, and is also affiliated with the Harris School of Public Policy. His research interest includes natural language processing, human-centered AI, and computational social science. Additional information is available at: https://chenhaot.com

6 Technical Requirements

For the in-person tutorial, we request roundtables so that participants can discuss together during the Q&A and the workshop activities; it would be good to have power outlets around the tables.

7 Ethics Statement

Our tutorial takes a human-centered perspective. We hope that our tutorial will broaden the scope of evaluations in NLP by introducing perspectives from HCI and psychology. This may help alleviate ethical concerns of NLP models in the long run by incorporating human perspectives into the development and evaluation process.

8 Special Themes

Our tutorial is alighed with the special theme of NAACL 2022, human-centered natural language processing.

References

Sara Aronowitz and Tania Lombrozo. 2020. Experiential explanation. *Topics in Cognitive Science*, 12(4):1321–1336.

Ivan A Aslanov, Yulia V Sudorgina, and Alexey A Kotov. 2021. The explanatory effect of a label: Its influence on a category persists even if we forget the label. *Frontiers in Psychology*, 12:745586–745586.

Jacob Bishop and Matthew A Verleger. 2013. The flipped classroom: A survey of the research. In *2013 ASEE Annual Conference & Exposition*, 10.18260/1-2–22585, Atlanta, Georgia. ASEE Conferences. Https://peer.asee.org/22585.

Thomas Blanchard, Nadya Vasilyeva, and Tania Lombrozo. 2018. Stability, breadth and guidance. *Philosophical Studies*, 175(9):2263–2283.

Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 454–464.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and Characterizing Human Rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307, Online. Association for Computational Linguistics.

Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2019. ERASER: A Benchmark to Evaluate Rationalized NLP Models. *arXiv preprint*. ArXiv: 1911.03429.

Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 275–285.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult.

Cristina Garbacea and Qiaozhu Mei. 2020. Neural Language Generation: Formulation, Methods, and Evaluation. *arXiv:2007.15780 [cs]*. ArXiv: 2007.15780.

Carly Giffin, Daniel Wilkenfeld, and Tania Lombrozo. 2017. The explanatory effect of a label: Explanations with named categories are more satisfying. *Cognition*, 168:357–369.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

Karthik S Gurumoorthy, Amit Dhurandhar, Guillermo Cecchi, and Charu Aggarwal. 2019. Efficient data representation by selecting prototypes with importance weights. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 260–269. IEEE.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language? *arXiv:2010.04119 [cs]*. ArXiv: 2010.04119.

Babak Hemmatian and Steven A Sloman. 2018. Community appeal: Explanation without information. *Journal of Experimental Psychology: General*, 147(11):1677.

Yuening Hu, Jordan Boyd-Graber, and Brianna Satinoff. 2011. Interactive topic modeling. In *Proceedings of the Association for Computational Linguistics*.

Abigail Z Jacobs, Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. The meaning and measurement of bias: lessons from natural language processing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 706–706.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14.

Frank C Keil. 2003. Folkscience: Coarse interpretations of a complex reality. *Trends in cognitive sciences*, 7(8):368–373.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Deanna Kuhn. 2001. How do people know? *Psychological science*, 12(1):1–8.

Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: A survey of empirical studies. *arXiv preprint arXiv:2112.11471*.

Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of FAT**.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*.

Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266.

Tania Lombrozo. 2012. Explanation and abductive inference.

Tania Lombrozo. 2016. Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, 20(10):748–759.

Tania Lombrozo, Daniel Wilkenfeld, T Lombrozo, and D Wilkenfeld. 2019. Mechanistic versus functional understanding. *Varieties of understanding: New perspectives from philosophy, psychology, and theology*, pages 209–229.

Jeffrey Lund, Piper Armstrong, Wilson Fearn, Stephen Cowley, Courtni Byun, Jordan Boyd-Graber, and Kevin Seppi. 2019. Automatic and human evaluation of local topic quality. In *Association for Computational Linguistics*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777.

D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research*, 13(3):334–359.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865.

Jake Snell, Kevin Swersky, and Richard S Zemel. 2017. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*.

Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. 2021. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for Fact Extraction and VERification. *arXiv:1803.05355 [cs]*. ArXiv: 1803.05355.

Sahil Verma, John Dickerson, and Keegan Hines. 2020. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*.

Eric Wallace, Matt Gardner, and Sameer Singh. 2020. Interpreting predictions of NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 20–23, Online. Association for Computational Linguistics.

Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328.

Sarah Wiegreffe and Ana Marasović. 2021. Teach Me to Explain: A Review of Datasets for Explainable NLP. *arXiv:2102.12060 [cs]*. ArXiv: 2102.12060.

Robert A Wilson and Frank Keil. 1998. The shadows and shallows of explanation. *Minds and machines*, 8(1):137–159.

James Woodward and Lauren Ross. 2021. Scientific Explanation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Summer 2021 edition. Metaphysics Research Lab, Stanford University.

Xi Ye, Rohan Nair, and Greg Durrett. 2021. Connecting attributions and qa model behavior on realistic counterfactuals. *arXiv preprint arXiv:2104.04515*.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "Annotator Rationales" to Improve Machine Learning for Text Categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305.