What Human Rationales Tell Us about Machine Explanations

Indira Sen[⋆], **Mattia Samory**[⋆], **Fabian Flöck**[⋆], **Claudia Wagner**[⋆†]
[⋆]GESIS, Cologne, Germany
[†]RWTH, Aachen
firstname.lastname@gesis.org

Abstract

It is important that NLP systems and their users find a common language to communicate which surface-level cues in a text are informative, be it to explain how the system performed an inference or how it should have. One way to evaluate machine explanations is to compute their overlap with human rationales. However, for which users and uses of NLP systems is this evaluation suitable? We discuss the factors which affect the consistency, validity, and reliability of human rationales as ground truth for machine explanations. Inspired by human-centered design, we build on these observations and outline directions to incorporate human rationales into the design process of NLP systems and to tailor them to the goals of different stakeholders of the system.

1 Introduction

NLP systems are widely used and interpetability is a highly desired property. One paradigm of evaluating interpretability is to compare local machine explanations—spans in a text that primarily informed model inference—against those that humans would deem most informative, also known as human rationales (DeYoung et al., 2020; Atanasova et al., 2020). This is in line with the longstanding practice of assuming humans as ground truth: as human-generated labels are ground truth for predictions, human-generated rationales should also be ground truth for machine explanations.[1]

While this an intuitive assumption in many cases, it is not without complications. First, the assumption may not hold true. Many tasks of interest in NLP, such as sexism and stance detection, aim at inferring subjective, ambiguous, or contested concepts, where humans themselves are not consistent or reliable (Gallie, 1955). Furthermore, several factors in how human rationales are defined, collected, and compared to machine explanations may affect their usefulness. For example, should all tokens in a human rationale have the same importance?

Inspired by calls for interdisciplinary research in building better NLP systems (Blodgett et al., 2020; Havens et al., 2020) and research in cognitive psychology and human-centered design, we reflect on the social and interactive functions of human reasoning, as a mode of *justifying oneself* and *convincing others* (Mercier and Sperber, 2017). This provides a lens for understanding the benefits and limitations of human rationales as ground truth for machine explanations. In particular, this allows connecting the goals of human rationales with those of different stakeholders of NLP systems (Norman, 1986; Suresh et al., 2021).

To elucidate the current gap in the use of human rationales in interpretability research, we first summarize how they were devised and are currently used in Section 2.3. Next, we enumerate the limitations of using human rationales for evaluating automated rationales, and identify open questions (Section 3). In Section 4 we discuss the potential utility of human rationales for explaining and augmenting NLP systems for different stakeholders.

2 Current Uses of Human Rationales

The recent resurgence of human rationales in the NLP community has been due to a growing interest in interpretability. We briefly discuss evaluations of NLP interpretability, the use of human rationales and the potential disconnect between the two.

2.1 Evaluating Machine Explanations

Machine explanations are often used to accomplish one of two goals: *plausibility* and *simulatibility*. For plausibility, evaluations of machine explanations include comparison with human rationales,

[1]We use the terms 'explanation' and 'rationale' interchangeably in this work.

a metric which can shed light on the convergence between human and machine reasoning. Rather than affording understanding how well humans understand how the machine derives the output from the input, the overlap of human rationales and machine explanations measure plausibility, or convincing a machine explanation would be to another human (Jacovi and Goldberg, 2020).

Besides plausibility, human-grounded simulatability experiments can be performed in lieu of application-oriented evaluations (Doshi-Velez and Kim, 2017). Simulatability evaluations, like forward simulation and counterfactual simulation, allows evaluating how well humans can predict model behavior on new inputs. This property is especially useful since it indicates that a person understands why an automated model predicts a certain label. These evaluations are based on explanations of predicted output rather than the ground truth since good explanations do not necessarily need to match human intuition.

There is no one-size-fits-all evaluation that works for all ML systems, even for the subset of ML-based NLP systems. Instead, evaluations should be tailored based on the goals of the system. For example, a classifier detecting moral foundations might have automated explanations with lower overlap with human rationales compared to a classifier detecting sentiment, simply because moral foundation is a much more subjective or *contested* construct. Furthermore, evaluation paradigms (such as overlap with human rationales, forward simulation) are aimed at diagnosing different aspects of interpretability, and the aspect being probed should be clearly defined to understand the quality of automated explanations. Finally, other goals of interpretability might also be considered such as mental model soundness and completeness (Kulesza et al., 2013).

2.2 The Use of Human Rationales

Human rationales have been solicited for several reasons. First, as a complementary source of supervision for machine learning models, i.e., to make machine learning models focus on the core features of the construct (Zaidan et al., 2007). The sole goal of the ML model is still to predict the class label in this setting. Second, they are also solicited to identify the spans where the construct features are salient in order to train ML models which could be used to predict these spans, rather than just the class label. See, e.g., the SemEval 2021 Toxic Spans Task.[2] The final and more recent use case of human rationales is as ground truth for evaluating machine explanations, for example as a part of the ERASER benchmark, for several NLP tasks (DeYoung et al., 2020).

2.3 The Disconnect Between Human Rationales and Machine Explanations

Lipton envisions several desiderata of model interpretability such as transferability, trust, and fair decision making. The implications of high and low overlap are unclear regarding the desiderata of model interpretability. Moreover, while measuring overlap with human rationales may provide an understanding of automated explanation quality for certain tasks, they might be misleading for others. For example, measuring overlap between human and machine rationales can help understand whether the predictions of spellcheck should be accepted or not, but the same measure cannot tell moderators much about subjective or ambiguous concepts like sexism or hate speech.

The cases of human rationales used as additional supervision, or as ground truth for machine explanations, brings forth a second disconnect. These use cases assume that human rationales contain all and only the patterns relevant for a task, guided by human commonsense knowledge, expertise, and symbolic reasoning abilities. Yet, NLP systems are also used to discover new knowledge, patterns, or theory from data. Indeed, increasingly machine learning systems outperform human experts in certain domains (Litjens et al., 2016; Schrittwieser et al., 2020), and may rely on explanations that are unaccounted for, or at least neglected by humans. Explanations of such systems need not overlap with human intuition, simply because the goal of the exercise is to discover hitherto unknown patterns. Therefore, both simulatability and overlap with human rationales are inappropriate evaluations of such use cases. **Therefore, there is a need for determining which NLP tasks can be evaluated based on human and machine rationale overlap.**

3 Limitations of Human Rationales for Evaluating Machine Explanations

In addition to the disconnect between the evaluation protocol and its purpose, we discuss other

[2] https://sites.google.com/view/toxicspans

issues and open questions associated with soliciting and analysing human rationales, including their purported status as ground truth for machine rationales.

3.1 Centering the Human in Human Rationales

Whose Rationales? Using human (annotator) rationales for evaluating automated rationales, assumes humans to be the ground truth, and that automated methods should mimic their reasoning. Yet, there are several drawbacks of this assumption.

We note that annotators label and reason based on their lived experiences, and their annotations are interpretive (Paullada et al., 2020). Indeed, previous research has shown that annotator disagreement cannot be dismissed as noise (Pavlick and Kwiatkowski, 2019; Gordon et al., 2021). This is particularly important for subjective concepts where annotator agreement can be low. Low agreement instances are often removed from training datasets, yet they may be the examples which provide a holistic understanding of the concept (Kenyon-Dean et al., 2018). Similarly, it is imperative to account for diversity of annotator rationales and find ways to aggregate them without losing their individuality.

Are Human Rationales Robust? It is entirely possible some explanations are non-spurious with respect to the current language use and construct specification (i.e., regardless of model, dataset, and annotators), that is they understand the correct association between the manifestation of the construct in text and it's latent factors. For example, the rationale for "women are terrible drivers" being 'women' and 'terrible drivers' makes the case of sexism due to the stereotype about women's bad driving. Other rationales may be spurious, for example, just 'terrible' in the previous sentence. While this rationale correctly spots one aspect of what makes this document sexist, it fails to account for the generalization of 'women'. How should the task of annotating rationales be designed such that they down-weight spurious features?

3.2 Validity of the Overlap

Lack of Consistency in Soliciting Human Rationales. There is no standard template for soliciting human rationales; some tasks asks for *comprehensive* rationales or *all* tokens that justify a decision, while others ask for *sufficient* tokens (DeYoung et al., 2020; Carton et al., 2020). Furthermore, the exact wording of the questionnaires is unknown. Given that seemingly innocuous changes in question wording can affect respondents' perception (Gendall and Hoek, 1990), it is important that we hone in on a standard template which can be customized based on different NLP tasks or use cases. Furthermore, there is a need for quality control in soliciting human rationales. Would the same human give the same explanation twice? How would one aggregate rationales across annotators and measure confidence or errors?

Challenges in Computing Overlap. The ERASER benchmark (DeYoung et al., 2020) describes two metrics for measuring overlap between annotator and machine rationales; discrete and soft selection. Since complete matching might penalize the inclusion or exclusion of trivial tokens in the explanation, a more relaxed metric which counts partial matches up to a certain threshold is introduced. Yet, the use of this metric begs the question of whether all tokens in an annotator rationale be ranked equal? For example, if an annotator justifies the positivity of the sentence "He's not too bad", based on the tokens 'not', 'too' and 'bad', two machine rationales with partial matches— 'not' 'bad' and 'too' 'bad', would be considered to have the same partial overlap with the annotator rationale. But it is clear that the second explanation is worse since it excludes the negation.

Granularity of Rationales We consider two common forms of human rationales— highlighted spans or tokens, or free text responses.[3] The former has several advantages but there are some open questions. Namely, how does one integrate background knowledge and commonsense reasoning when rationales are lacking (e.g. because the explanation is indirectly related to the construct)? Next, how to integrate latent rationales?

When Predictions are Wrong. Current research computing overlap between human and machine rationales tend to ignore wrong predictions. What are the implications of overlap between human and machine rationales, even for misclassifications? There is an outstanding need to explore the interaction between local explanations, human rationales, and predictions.

There is no one-size-fits-all evaluation that

[3]There is a third emergent form of human rationales for NLP tasks, namely 'structured' responses (Wiegreffe and Marasović, 2021). These responses are not entirely free-form but rather in response to template-like questions. Given their recency, we hope to study structured rationales in future work.

works for all NLP systems. Instead, **evaluations should be tailored based on the goal of the system.** A classifier detecting moral values might have automated explanations with lower overlap with human rationales compared to a classifier detecting binary sentiment in reviews, simply because moral value is a more subjective concept *and* is harder to anchor to specific tokens than sentiment.

4 The Potentials of Human Rationales

We now turn to how annotator rationales, in the form of either in-text spans or free text, *might* help in facilitating machine intelligence, rather than evaluating it. by opting for a human-centered design lens (Norman, 1986). Human-centred design seeks to include all stakeholders of a potential system and their thoughts, goals, and needs in the design process. To that end, we describe how annotator rationales might impact on and interact with three types of stakeholders—annotators, developers of the NLP system, and users of the NLP system.

4.1 Annotators

Annotators should be considered an important stakeholder in the design of NLP systems and their input, an important part of the design process. In fact, annotators are essential in the current paradigm of supervised ML-based NLP systems. One can draw parallels between annotator rationales and **design rationales**, which explicitly list decisions made during a annotation process, and the reasons behind them (Moran and Carroll, 2020).

The process of annotation can incorporate human-centred approaches which allow annotators to provide greater insight and justification into their reasoning, for example through the think aloud protocol. Drawing from the social and persuasive function of human reasoning (Mercier and Sperber, 2017; Pruthi et al., 2020), annotator rationales can also be an example of inter-annotator communication or deliberative crowdsourcing (Schaekermann et al., 2018; Tang et al., 2019). To that end, annotator rationales can improve consensus building and dispute resolution, the latter being a regular fixture in qualitative coding. Finally, incorporating annotator rationales more deeply into the design process of NLP systems also seeks to further recognize and value the contribution of annotators.

4.2 Model Developers

Model developers can use annotator rationales as a secondary source of supervision (Zaidan et al., 2007). More importantly, annotator rationales can inform the design process of the model (for example during data cleaning or feature engineering) even before the operationalization process, or while defining the construct that they aim to measure. Annotator rationales can reveal blindspots in construct definition and operationalization.

For subjective or contested concepts, annotator rationales can help model developers in several ways; first, they provide an understanding of the design space of the construct, specifically, the ambiguous parts that lead to dispute. Second, model developers can estimate an upper limit of the predictability of the construct, and accordingly reflect on the repercussions of deploying the system.

4.3 Model Users

We think of two types of model users— decision makers and sense makers. Decision makers, for example, content moderators or recruiters using the output of NLP systems to guide their judgement, might want to **measure how well a model replicates the human evidence-based decision-making process.** Here, decision makers can compare automated rationales and human rationales but with more nuanced metrics that take into account the weight of different tokens. Annotator rationales also indirectly affect decision makers by providing information on whose lived experiences are encoded in the model, and by scaffolding their expectations from the model.

Sense makers, such as researchers in humanities or social science using NLP systems to find patterns in text, might prefer explanations that facilitate **breaking down complexity to find surprising associations** similar to a grounded theory approach (Nguyen, 2018). Annotator rationales can help sense makers by serving as a starting point of theories which can augment the process of discovery in an abductive fashion (Walton, 2014).

To conclude, annotator rationales can be a powerful source of information across the entire design process, **providing insight on why people, specifically annotators, perceive constructs the way they do, unearthing considerations outside of the model designer's purview, and indirectly increasing model users' trust.**

There could very well be other stakeholders, par-

ticularly targets of NLP system predictions. We acknowledge that they are an important stakeholder, especially since they might be disproportionately harmed. While annotator rationales do not directly affect them, we hope to understand the values and needs of targets of NLP systems in more depth in future work.

5 Discussion

We enumerate several issues and open questions about the use of annotator rationales in evaluating machine rationales. We hope that the questions posed here will help us, as a community, reflect on the implications of this type of evaluation, while improving it's standardization and specification, as well as discerning use cases for which such an evaluation is informative. We suggest that the design of evaluations for explanations keep in mind the needs and goals of different stakeholders. For example, for our envisioned stakeholders, while simulatability and overlap with annotator rationales might help model developers, it is unclear if they benefit sense makers. Finally, we layout several suggestions and connections to literature in HCI that may reveal interesting uses of human rationales in unearthing how humans individually and collectively reason, thereby allowing us to incorporate this reasoning into NLP systems. In future, we hope to build upon the stakeholders and use cases described in this work, to design better NLP systems, their explanations, and the evaluation of these explanations.

Acknowledgements. We thank the anonymous reviewers of the HCI+NLP workshop for their constructive feedback and helpful pointers.

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

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.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

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

Walter Bryce Gallie. 1955. Essentially contested concepts. In *Proceedings of the Aristotelian society*, volume 56, pages 167–198. JSTOR.

Philip Gendall and Janet Hoek. 1990. A question of wording. *Marketing Bulletin*, 1(5):25–36.

Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality.

Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2020. Situated data, situated systems: A methodology to engage with power relations in natural language processing research. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 107–124, Barcelona, Spain (Online). Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, et al. 2018. Sentiment analysis: It's complicated! 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 1886–1895.

Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*, pages 3–10. IEEE.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Geert Litjens, Clara I Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen-Van De Kaa, Peter Bult, Bram Van Ginneken, and Jeroen Van Der Laak. 2016. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports*, 6(1):1–11.

Hugo Mercier and Dan Sperber. 2017. *The enigma of reason*. Harvard University Press.

Thomas P Moran and John M Carroll. 2020. *Design rationale: Concepts, techniques, and use*. CRC Press.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. 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 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

Donald Norman. 1986. User centered system design. *New perspectives on human-computer interaction*.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2020. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv preprint arXiv:2012.05345*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Danish Pruthi, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C Lipton, Graham Neubig, and William W Cohen. 2020. Evaluating explanations: How much do explanations from the teacher aid students? *arXiv preprint arXiv:2012.00893*.

Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–19.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. 2020. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609.

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. *arXiv preprint arXiv:2101.09824*.

Wei Tang, Ming Yin, and Chien-Ju Ho. 2019. Leveraging peer communication to enhance crowdsourcing. In *The World Wide Web Conference*, pages 1794–1805.

Douglas Walton. 2014. *Abductive reasoning*. University of Alabama Press.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable nlp. *arXiv preprint arXiv:2102.12060*.

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.