"Get Their Hands Dirty, Not Mine": On Researcher-Annotator Collaboration and the Agency of Annotators

Shengqi Zhu
Cornell University
sz595@cornell.edu

Jeffrey M. Rzeszotarski
Cornell University
jeffrz@cornell.edu

Abstract

Annotation quality is often framed as post-hoc cleanup of issues caused by annotators. This position paper discusses *how* and *why* this narrative limits the scope of improving data quality. We call to consider annotation as a procedural collaboration, outlining three key points in this direction: (1) An issue can be either *annotator-* or *researcher-oriented*, where one party is accountable and the other party may lack ability to fix it; (2) yet, they can co-occur and/or have similar consequences, and any specific problem we observe may be a complicated combination; (3) therefore, we need a new language to capture the nuance and holistically describe the full procedure to resolve these issues. To that end, we propose to study how *agency* is manifested in annotation and picture how this perspective benefits the community more broadly.

1 Introduction

The pursuit of data quality in NLP and AI usually takes a post-hoc, outcome-oriented approach, targeting the immediate goal of refining the data obtained. This includes the vast amount of ML-based post-processing (Raykar et al., 2009; Khetan and Oh, 2016), and recent notions for replacing crowd work (Gilardi et al., 2023; Alizadeh et al., 2023). While more works recently seek to dig deeper, attention is still largely placed on retrospectively analyzing artifacts in existing datasets (Malaviya et al., 2022; Gururangan et al., 2018; Sap et al., 2022).

Emphasizing the output of a reliable set of annotations over the full data collection procedure often leads data issues to be attributed as undesired *noises* or *artifacts* from *annotators' behaviors, biases, or failure* (Rodrigues and Pereira, 2018; Han et al., 2020, etc.) With this narrative, annotators are implicitly stationed as the source of problems, and researchers as the party who then find, correct, and/or discard the flawed data. This could lead to an oversight of various systematic failures that attributed more to researchers, e.g., a flaw in the early stage of task design (Gururangan et al., 2018; Pyatkin et al., 2023; Gadiraju et al., 2017, etc.)

Complicating this further, these independent issues – with potentially distinct sources and solutions – often co-occur, interact, and converge in forming the eventual, real-world problems we encounter. This creates a subtle chasm between theory and practice: while various independent issues are well-documented, annotation is still often understood only via retrospective observations. On the one hand, many works discuss the extensive intricacies of annotation tasks as well as what researchers should look for; On the other hand, improving data quality in practice remains a whac-a-mole game: How data were collected doesn't seem relevant until something pops up; we then guesstimate what happened way back and attempt a fix, and again use it as normal until the next issue stands out. As users downstream are agnostic of the data collection details, this mode is limited.

It is especially hard in retrospective analysis, if users are motivated at all, to further search through the long list of known issues and narrow down to some exact match in the literature with their guesstimates. This calls for a universal language that systematically captures how issues arise and interact in annotation pipelines, one that composes and compares the practices of the researchers and annotators in a coherent, generalizable framework.

In this position paper, we urge the community to rethink the acquisition of annotations as a bilateral, collaborative process: the researchers as *task designers* and the annotators as *task accomplishers*, where responsibility is shared among both parties. To better describe flaws in such procedures, we propose to study the *agency* of annotators, i.e., how much capacity to act is allocated by task designers to an annotator toward the best outcome. We set forth an essential **agency (mis)alignment problem**: while researchers *intend* to provide a certain

Findings of the Association for Computational Linguistics: ACL 2024, pages 8773–8782

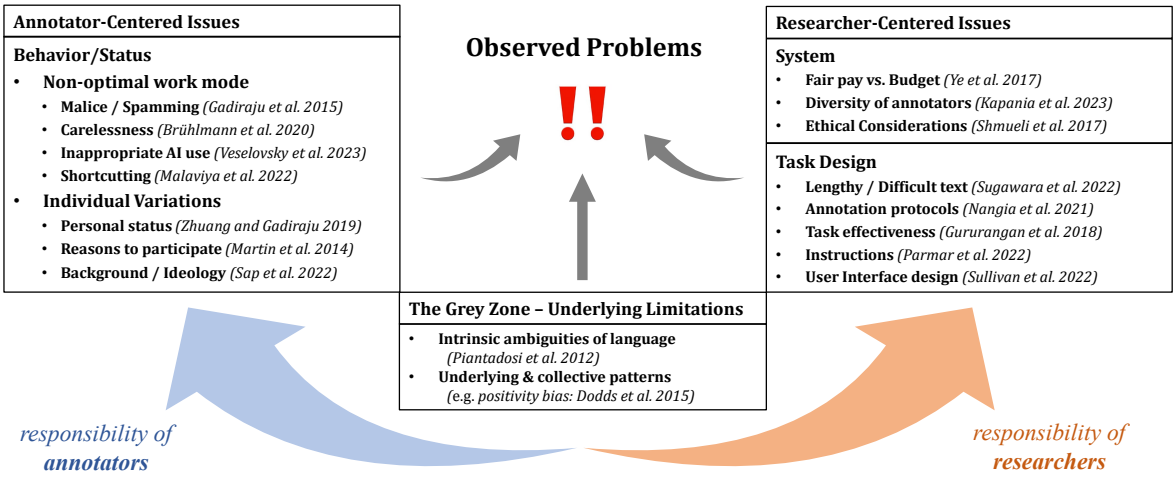

Figure 1: The Spectrum of Accountability. We extract and categorize atomic issues in the existing literature, grouped as researcher- or annotator-centered on the two ends of the spectrum. Additionally, we mark a "grey zone" caused by inherent limitations in human language and cognition, which does not notably attribute to any single party. Each issue is displayed in pair with one representative piece of work in related fields, and we discuss the issues in more detail in Appendix A for interested readers to initiate a literature search. The atomic issues along the spectrum contribute together and interact in complex ways to form the problems we encounter in practice.

degree of annotator agency, the task setup they *implement* may entail a different level of agency, leading to undesired outcomes. With this perspective, we can better understand the annotation pipeline with a procedural lens and match downstream observations with their root causes. We hope this will aid in the broader mission to empower data work with additional transparency and accountability.

2 Tracking down the "Problems" We See

Our statement is based upon several pioneering pieces that consider the annotation process beyond annotators' behaviors and their outcomes. Parmar et al. (2023) were among the first to point out that existing work targets *annotator-related bias*; instead, they measure *instruction bias*: artifacts introduced as early as in the instructions, which directly leads to downstream biases. Huang et al. (2023) surveys workers' perspectives and how they view their assigned roles. They show that the clarity of tasks impacts workers' willingness and effort level needed to complete them. Plank (2022) review the *human label variation* "problem": disagreements in annotation may not always be noise, but instead suggest important properties of the data and task (e.g. hard cases). Finally, Rottger et al. (2022) highlight that the definition of annotation is essentially different within two contrasting paradigms, "descriptive" and "prescriptive" annotation, based on whether subjectivity should be encouraged.

These individual pieces share a common insight: implementing a task with crowd work is only a fraction of the data-annotation pipeline. While this phase has many of its own flaws, it also reflects subtleties from earlier stages such as how researchers plan out and describe their tasks.

Consider two hypothetical failure cases for a research group seeking annotations that "represent a good output from a chatbot". In one case, the researchers used this generic description in quotes as the sole instruction, and the annotations they got were full of artifacts and inconsistencies. In the other case, several annotators improperly used ChatGPT when they were clearly advised not to, and thus the results are unusable. For annotators in the first case, the best work would not possibly exceed a guesstimate of the researchers' intention: the "chatbot" in question and what would be a "good output" is agnostic. However, things could be largely different if the *researchers* take a moment to encode the fundamentals of their goal in the task design. Conversely, in the second case, as long as some *annotators* insist on the improper move, a researcher won't be able to easily get rid of the trouble of extra work and wasted funds. While both can be reported as a quality issue, the actual cause and solution are distinct: one centering on the role of the researchers and the other on the annotators.

Compounded observations and atomic issues To add to the above, any observed problem in the real world can often be a compound of various contributors: a lack of motivation found among an-

notators may well be individual negligence (annotator), overly low payment (researcher), or – more often – a subtle mixture in the middle. The annotation procedure essentially follows the law of the minimum: any part going wrong would be sufficient to undermine the outcome. An oversight early in the instruction or protocol similarly determines the lower bound of annotation quality, but tackling the annotators' side in these cases has little effect. The borders are further blurred as issues centering different roles can have similar appearances. For instance, in the first failure case above, an annotator may misunderstand the task and resort to ChatGPT with no bad intention but just because they believe that's how they should find "good chatbot outputs".

To sort out the complexity of real-world observations, we outline *atomic* issues (which derive from a single root cause) in Figure 1, categorized by *accountability*, i.e., which party would be the center cause for the issue, and further detail them in Appendix A to link the related literature for interested readers. These issues on the spectrum can interact and even converge, forming the *compounded* problems we encounter when retrospectively examining the efficacy of annotations.

Note that there are no fewer *researcher-centered* issues than the commonly known *annotator-centered* ones. Moreover, the inherently asymmetric relationship between researchers and annotators can create a risky situation: deficiencies on the former party may have a high impact, but are masked by the fact that annotators are the ones who perform the work. For instance, breakdowns in a strict manual might be more likely attributed to annotator laziness than a failure to properly deliver instructions. While issues are well discussed individually in the literature, understanding real-world fallacies demands a holistic model that involves all these subtleties throughout the pipeline. We aim to draw attention to the complexity of the interplay between these issues in reality, and further propose a method to describe it in the next paragraph.

The Agency Problem Researcher-centered issues may appear similar to annotator-centered ones, but their solutions are distinct. How should such intricacies be collected and modeled in any specific task? When task designers *conceive* a task, it might seem at first blush that they take control of all that needs to be controlled. However, tasks when *implemented* provide varying (and sometimes unexpected) degrees of *agency* to annotators in the course of their work. For example, meticulous manuals might be deployed to confine annotators' operations in a labeling task, but there can be unforeseen or underspecified interaction modes in how the rules are deployed by annotators. In other words, there remains extra agency for annotators themselves to determine how to complete their tasks. This therefore leads to *annotator-centered issues* not being monitored by researchers.

We call attention to such *agency misalignment problem* in annotation workflows. On one hand, we have the notion of *annotator's agency*: the extent to which an annotator is able to take (various) actions towards better annotation quality. Constructing any dataset demands a certain level of annotator's agency; yet, a capable annotator's agency is also the direct result of how a task is deployed by its designers. This leads to the other component, a *task designer's intended agency*. Task designers conceive a task with an enhanced or constrained space of possible actions, along with the expectations that these actions will be taken (e.g., the strict manual in the previous example). They also might instead promote agency on the part of annotators to gather more heterogeneous data, such as in the case of chatbots. Nonetheless, these conceptions might not be fully achieved, thus creating *misalignments* between the conception and implementation.

We pose that misalignments between intended and realized agency are not only a major contributing factor in quality issues, but studying it also helps to organize and root out these issues. For instance, in the previous codebook example, misalignments between the designers' intended low-agency task and the actual freedom annotators had to interpret and apply the codebook led to a series of breakdowns that reduced data quality. Yet, without this integrated view of the process, it is easy to attribute this breakdown solely to unruly annotators. This might lead a designer to attempt to recruit "better" workers, when an effective solution is instead to revise the task design.

3 Case Study: A tale of 4 datasets

To further our discussion of the agency misalignment problem, we will outline four classic dataset case studies, visualized as a detailed diagram in Figure 2. Each dataset represents a branch of NLP tasks it belongs to, but also features a specific implementation in question.

Consider ① the Part-of-Speech (POS) tagging

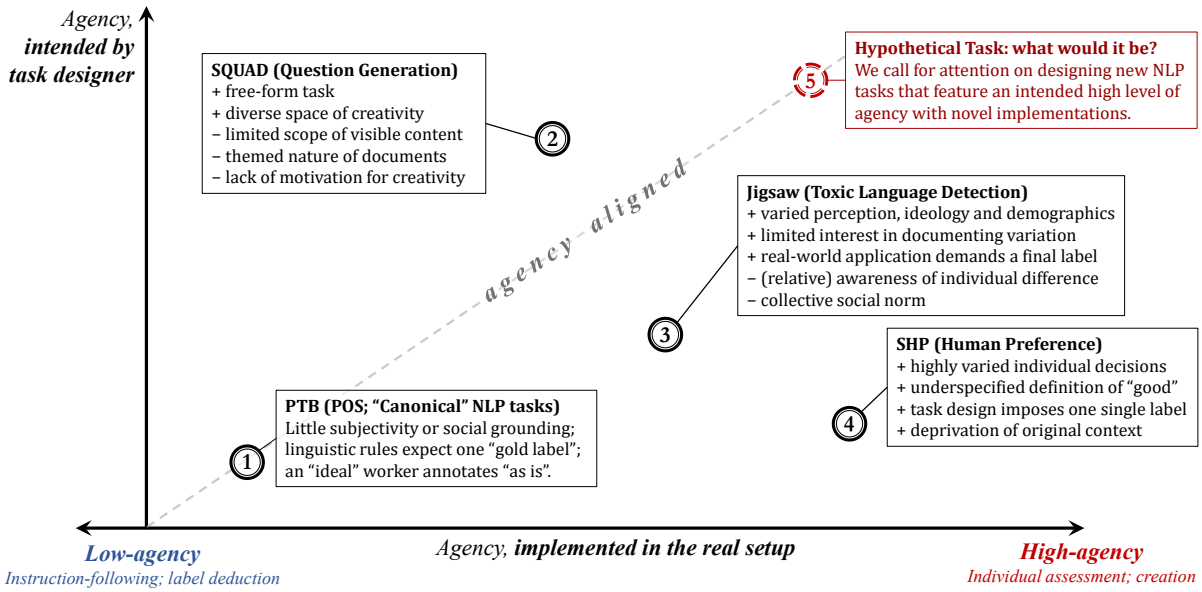

Figure 2: A diagram of agency (mis)alignment over four widely-used benchmarks: Penn TreeBank, SQUAD, Jigsaw, and SHP. For each dataset, a "+" or "-" sign indicates a factor that causes the annotators' agency to increase or decrease respectively in the *implemented* annotation procedure. Factors can be either intentional design (e.g., SQUAD being free-form) or limitations (e.g., SHP simplifying Reddit upvotes as binary choices); if a dataset owns multiple factors toward higher (or lower) agency but is not *intended* to be so, a misalignment occurs. (It would also be interesting to explore if we can design tasks that both intend and implement high agency.)

setup of the Penn TreeBank (PTB, Marcus et al. 1993) as a base. Representing the "canonical" linguistic annotation setup, the task is rigidly grounded with linguistic definitions mapping each word ("*the*") to a label ("*DET*") on a largely one-to-one basis. Annotators are systematically informed with a coherent definition and variance is discouraged. Both the researchers and annotators work under a shared set of rules to reduce irregularities. Thus, the intended and actual agency assigned are aligned at a minimal level[1], as denoted in the lower left, and the resulting data generally matches the expectations established at the start.

In contrast, creativity is essential for downstream NLP tasks like ② SQUAD (Rajpurkar et al., 2016). Making sure that annotators have agency is key to providing the necessary variability. To achieve this, a paragraph is shown and annotators are prompted to generate five questions and highlight text snippets as the corresponding answers. To boost the variance, the task takes a self-guided form with free-form input, and users are hinted to "use your own words" and "avoid the same words".

Nonetheless, the implementation did not deliver a full range of creative responses. While little constraints are imposed by the *instructions*, annotators are limited by the *source* of the text: all paragraphs come from the Introduction section of a Wikipedia item, which takes a highly similar form and revolves around a topic entity. Annotations, unsurprisingly, follow these patterns. Moreover, no actual elaborations or motivations are given to narrow down how (not) to use the same strategies *across* instances. Annotators, faced with similar sources that have similar patterns, may feel compelled to follow them. While they theoretically may have agency to answer freely (and are encouraged to do so), shortcut-taking and mimicry have been reported, essentially making the data "simpler" than conceived (Sugawara et al., 2018; Bartolo et al., 2020). In this case, a misalignment between expectations and the way the task format realized them led to problematic outcomes.

Underspecification of a task may also create issues. Consider the family of datasets regarding Toxic Language Detection like ③ Jigsaw's Perspective API. Intrinsically, there is no single definition for what language is "toxic". Annotators'

[1]Note that this is not to say an aligned, low-agency task always leads to unanimity; rather, it means the task is implemented in such a way that *a relatively small space of variation is allowed for any individual annotator*. In fact, "genuine disagreements" with multiple acceptable answers have been widely found in POS tagging (Plank et al., 2014). This maps to the "grey zone" on the spectrum (included as "Intrinsic ambiguities of language"): fundamental factors beyond the control of both parties also contribute to observed "problems".

perception is related to (if not determined by) their personalities, demographics, and ideologies (Sap et al., 2022). Even if a designer were to construct an extremely detailed guideline, there is no guarantee individual differences may not be a factor (more so due to subjectivity Waseem et al. 2021). Moreover, commercial platforms will eventually need to make an exact decision whether or not to intervene with or remove a potentially toxic speech, further constraining the decision space as a yes-or-no choice downstream. In this case, misalignment between an expected, rigid outcome and the messy realities of human judgment can challenge the ability to gather useful annotations.

This contrast of *varied individuals* vs. *one simplified decision* is even more polarized in Human Feedback datasets represented by ④ Stanford Human Feedback (Ethayarajh et al., 2022), which compares Reddit upvotes to entail an A-or-B preference decision. Here, *preference* was conceived as a task with very low agency: the only operation possible is the up/downvote made, and it does not vary by topic or time. However, real-life preferences and upvotes are highly arbitrary, and it is not clear whether the latter is a decent proxy of the former – for instance, upvotes can be given to arbitrarily many replies (high-agency) but are restricted to strictly one of the two in pairs (low-agency). Misaligning the diverse nature of the implementation with a low-agency framework risks misrepresenting the ground truth embedded in the annotations.

4 Discussion

Improving annotation quality is all about the interplay of annotators and designers. Our framing of the agency misalignment problem essentially proxies a larger misalignment between *what we think we're building* and *what we end up with*. Rather, anything *could* be part of *a* dataset; but the designers themselves should – and only are they able to – ensure it is *the* dataset, the exact one that we believe to serve a certain purpose. Given our data-hungry models, the stakes are high. A piece of data from a coarse design may instantly have a consequence on a sophisticated application: to determine which model is better, to place rewards on some options while penalizing others, and – eventually – to decide what values are coded in our models and whose input is represented. Shifting our focus and transferring the responsibilities in data collection also means letting go of control over our models, eventually adding to the already notorious myth of training data and evaluation (Paullada et al., 2021; Rogers, 2021; Howcroft et al., 2020).

We hope our proposed framing can be an entry point into understanding what our data are and how they are generated, as early as when we conceive it as a hypothetical task. We see two ways this framing might be used. First, as a diagnostic, examining issues of misalignment can help to explain the human factors underpinning what may seem like a simple issue of low-quality annotations. This can lead to fixes with greater impact across multiple task designs. Second, using this framing prospectively when designing a task can help to identify unforeseen complications as agency is allocated. For instance, thinking carefully about the degrees of freedom a codebook affords can help to identify its vulnerabilities to unusual annotator behavior.

In considering this issue, we also draw attention to discussions in the field of crowdsourcing related to working conditions (Kittur et al., 2013), techniques for shepherding crowd work (Dow et al., 2012), alternate collaborative annotation workflows (Chang et al., 2017), perspectivist judgments (Cabitza et al., 2023), as well as a wide range of quality control techniques under debate within the domain (Daniel et al., 2018; Rzeszotarski and Kittur, 2012; Zaidan and Callison-Burch, 2011).

5 Conclusion

We hope this position paper offers readers an additional perspective on how to approach annotation quality issues and their compounds via a cohesive and procedural lens. Moving forward, we urge the community to adopt more holistic assessments of annotation pipelines, and draw insights from the ways that human factors, crowdsourcing, and NLP researchers are innovating in how tasks are designed and assessed.

Limitations

This position paper emphasizes the procedural collaboration between researchers and annotators and initializes discussions on how that collaboration can be modeled. The work, as a succinct proposal, is necessarily limited for a rather broad shift from the deeply rooted discourse and ecosystem. The spectrum and case studies may not cover all related issues on the table, and subsequent work would be necessary to quantify the misalignment of agency and validate the expressiveness of such a model.

Nonetheless, we hope that our analysis will serve as an entry point for further discussion towards a better future of data collection, as well as a bridge connecting our field to the immense work in related fields navigating toward a shared mission.

Ethical Considerations

Data as the ingredients of models determines what decisions they make and what values are encoded. Our work is in line with this direction and aims to help understand and improve data collection for that reason. While our work does not contain new datasets or experiments with humans, we propose a new way to analyze the annotation procedure based on agency as defined. This may influence how data collection practices and the resulting datasets are described and assessed in the future. Yet, we believe these new perspectives in this position paper would enrich current views in a positive way.

Acknowledgements

We thank our reviewers and the members of the Artifact Lab @ Cornell InfoSci for their detailed feedback and suggestions on our drafts.

References

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks. *arXiv preprint arXiv:2307.02179*.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Florian Brühlmann, Serge Petralito, Lena F Aeschbach, and Klaus Opwis. 2020. The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology*, 2:100022.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2334–2346.

Paul G Curran. 2016. Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66:4–19.

Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1):1–40.

Peter Sheridan Dodds, Eric M Clark, Suma Desu, Morgan R Frank, Andrew J Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M Kloumann, James P Bagrow, et al. 2015. Human language reveals a universal positivity bias. *Proceedings of the national academy of sciences*, 112(8):2389–2394.

Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 1013–1022.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 1631–1640.

Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2017. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM conference on hypertext and social media*, pages 5–14.

Gerd Gigerenzer and Peter M Todd. 1999. *Simple heuristics that make us smart*. Oxford University Press, USA.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Donghoon Han, Juho Kim, and Alice Oh. 2020. Reducing annotation artifacts in crowdsourcing datasets

for natural language processing. In *The eighth AAAI Conference on Human Computation and Crowdsourcing. AAAI*.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Jeff Howe. 2009. *Crowdsourcing: Why the power of the crowd is driving the future of business*. Crown Currency.

Olivia Huang, Eve Fleisig, and Dan Klein. 2023. Incorporating worker perspectives into MTurk annotation practices for NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1010–1028, Singapore. Association for Computational Linguistics.

Quzhe Huang, Shibo Hao, Yuan Ye, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. 2022. Does recommend-revise produce reliable annotations? an analysis on missing instances in DocRED. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6241–6252, Dublin, Ireland. Association for Computational Linguistics.

Shivani Kapania, Alex S Taylor, and Ding Wang. 2023. A hunt for the snark: Annotator diversity in data practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Ashish Khetan and Sewoong Oh. 2016. Achieving budget-optimality with adaptive schemes in crowdsourcing. *Advances in Neural Information Processing Systems*, 29.

Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1301–1318.

Chaitanya Malaviya, Sudeep Bhatia, and Mark Yatskar. 2022. Cascading biases: Investigating the effect of heuristic annotation strategies on data and models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6525–6540, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

David Martin, Benjamin V Hanrahan, Jacki O'neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 224–235.

Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R. Bowman. 2021. What ingredients make for an effective crowdsourcing protocol for difficult NLU data collection tasks? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1221–1235, Online. Association for Computational Linguistics.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. Don't blame the annotator: Bias already starts in the annotation instructions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1779–1789, Dubrovnik, Croatia. Association for Computational Linguistics.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11).

Steven T Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Valentina Pyatkin, Frances Yung, Merel C. J. Scholman, Reut Tsarfaty, Ido Dagan, and Vera Demberg. 2023. Design choices for crowdsourcing implicit discourse relations: Revealing the biases introduced by task design. *Transactions of the Association for Computational Linguistics*, 11:1014–1032.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Vikas C Raykar, Shipeng Yu, Linda H Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. 2009. Supervised

learning from multiple experts: whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual international conference on machine learning*, pages 889–896.

Filipe Rodrigues and Francisco Pereira. 2018. Deep learning from crowds. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Anna Rogers. 2021. Changing the world by changing the data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2182–2194, Online. Association for Computational Linguistics.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Jeffrey Rzeszotarski and Aniket Kittur. 2012. Crowdscape: interactively visualizing user behavior and output. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pages 55–62.

Jeffrey M Rzeszotarski and Aniket Kittur. 2011. Instrumenting the crowd: using implicit behavioral measures to predict task performance. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 13–22.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, Online. Association for Computational Linguistics.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium. Association for Computational Linguistics.

Saku Sugawara, Nikita Nangia, Alex Warstadt, and Samuel Bowman. 2022. What makes reading comprehension questions difficult? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6951–6971, Dublin, Ireland. Association for Computational Linguistics.

Jamar Sullivan Jr., Will Brackenbury, Andrew McNutt, Kevin Bryson, Kwam Byll, Yuxin Chen, Michael Littman, Chenhao Tan, and Blase Ur. 2022. Explaining why: How instructions and user interfaces impact annotator rationales when labeling text data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 521–531, Seattle, United States. Association for Computational Linguistics.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*.

Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied machine learning: On the illusion of objectivity in nlp. *arXiv preprint arXiv:2101.11974*.

Teng Ye, Sangseok You, and Lionel Robert Jr. 2017. When does more money work? examining the role of perceived fairness in pay on the performance quality of crowdworkers. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 327–336.

Omar Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 1220–1229.

Mengdie Zhuang and Ujwal Gadiraju. 2019. In what mood are you today? an analysis of crowd workers' mood, performance and engagement. In *Proceedings of the 10th ACM Conference on Web Science*, pages 373–382.

A Details of the atomic issues along the Spectrum of Accountability

We provide more details and recent examples for the atomic issues that we identify and cluster along the Spectrum of Accountability (Fig. 1). We hope this could provide readers with the general landscape of the issues of interest – especially readers who are not specifically familiar with the field of annotation and crowdsourcing – and initiate future research and discussions. Note that the discussion

here for each issue is by no means an exhaustive review of that direction, but we seek to provide an appropriate starting point for interested readers to search into the broader related literature.

A.1 Annotator-Centered Issues

The usability of annotated data is directly determined by the work of annotators. For a specific task, annotators might have not performed their work effectively and properly (*None-optimal Work Mode*). More broadly, one's role of (outsourcing) workers might be compromised or varied due to individual conditions (*Individual Variations*).

Non-optimal work mode is a major source of issues where the annotation work is compromised as the result of annotators' (purposeful) behaviors for a specific task, even when seemingly qualified after adequate consideration of the researchers (e.g. via a pre-test or a filter). This would be especially relevant for crowdsourcing scenarios where task designers usually only have limited and distant supervision of the implementations ("outsourcing to an undefined, generally large group", as defined by Jeff Howe (2009)). These cases can be further categorized by whether and to what extent the undesired mode is intentional and destructive. On one extreme, it is practically impossible to locate a participant with a determined and strategic *malicious intent* (Gadiraju et al., 2015), since pre-task screening could be deluded in most cases. In a milder scenario, users may not intend to sabotage the requests, but (knowingly) react to the task without adequate care as instructed (Curran, 2016; Brühlmann et al., 2020).[2] Annotators can also strategically develop *shortcuts* that approximate the actual requirements and instructions with a simplified but inaccurate implementation (Malaviya et al., 2022).[3] These have since inspired a broad literature in quality control and estimation (Daniel et al., 2018; Rzeszotarski and Kittur, 2011; Zaidan and Callison-Burch, 2011). With the recent progress in the field of Artificial Intelligence, another issue is the improper use of AI tools. The extreme ease of access to proxies like large language models has proliferated this type of improper work mode (when they are not suitable for the task) (Veselovsky et al., 2023).

Individual Variations are another group of issues that are less specific to the implementation of any single task. Rather, they concerns the characteristics and roles of individual annotators on a higher level. For one thing, factors that are out of the reach of researchers, e.g., the mood (Zhuang and Gadiraju, 2019) when a worker starts on their task, may always have a direct impact on the output. There are also more fundamental aspects in the profile of annotators that subtly impact the way they work. One important aspect is the motivation or "reason" that an annotation worker is in their role (Martin et al., 2014). Crowd work may, for instance, merely be an option of entertainment over a cup of tea (that even pays!) for one retired scholar, but is the lifeline of another unemployed single mother. While both can be well qualified, the outcome can have different patterns due to personal status. In the former case, the annotator might be happy to provide an especially detailed and critical response with feedback, but would only take on an extremely small portion of the data, and has long intervals between their work periods. In the latter, the annotator may be willing to efficiently complete tasks in large batches, but the speed and incentive might be prioritized over higher quality, and they might be subject to the effects of fatigue. Similarly, the background and mindset of annotators can contribute to notable variations encoded in the outcome, which would be both a rich source of high-agency tasks and a major challenge to monitor and mitigate inappropriate biases (Sap et al., 2022).

A.2 Researcher-Centered Issues

Like the annotators' side, researchers can run into atomic issues both from the specific design choices in a task (*Task Design*) and from the broader considerations for annotations as a (societal) *system*.

System While data collection usually has the most to do with a research purpose, it should not be omitted that the researchers are the payers and employers of crowd work. Various issues can emerge within the operation of this system with unequal roles. The most classic one is perhaps the topic of

[2] Any human worker would have a natural error rate and this is not the major referent in this context. Instead, we refer to the cases where we see a systematic pattern of carelessness in one's work.

[3] Malaviya et al. 2022 refer to these broadly as *cognitive heuristics* (Tversky and Kahneman, 1974; Gigerenzer and Todd, 1999), yet majorly concerns the deliberate application of these strategies. We avoid the subtlety introduced by the vast literature in the Psychology field surrounding this terminology, as we seek to make a clear, three-way distinction that corresponds to different accountabilities: the strategic, intentional shortcutting of annotators (*Shortcutting* in the spectrum), the intuitive, less-aware cognitive patterns (*Underlying & collective patterns*), and an act of expedience faced with unclear instructions or flawed task design.

fair pay and, as a result, the perceived fairness and incentives of workers (not to forget the researchers' own budgets are in the play as well [Ye et al., 2017; Huang et al., 2023, among many others]). Recent studies also pose questions on the broader ethics in annotation and crowdsourcing performed by workers (Shmueli et al., 2021). In the role of employers, researchers are also in charge of the selection and filtering of participating workers. This complements the improper work mode of annotators: researchers should, as the upstream designer and employer, guarantee the effectiveness of their data by considering who would be suitable workers, before attributing to individual failures.

Task Design The design of a task is the very core of its eventual usability. An overlooked design factor might directly lead to the failure of communication with workers and undetected deviations from the supposed scenario (as our discussions of the agency are concerned). Various hierarchies work together to form a valid task and the staging could well be a whole separate paper on its own. Things to consider involve the abstract, early-stage conception of protocols and procedures to use (Nangia et al., 2021), whether a task-specific protocol will work out in practice (Huang et al., 2022), as well as how effective this task itself is effective and complete at all (e.g., the classic findings of Gururangan et al. (2018) on the strong yet overlooked cues in Natural Language Inference). Meanwhile, it also spans to the very details when handing over to the annotation workers, such as whether the text layout has been frustratingly long or difficult (Sugawara et al., 2022) and how the annotation User Interface should be displayed (Sullivan Jr. et al., 2022).

A.3 The "Grey Zone" – Underlying Limitations

Finally, we would also like to highlight the more fundamental and universal sources of issues which we call *the grey zone*. Intricate as the cognitive system and language ability of humans are, many times a failure in a task does not simply go to the party that designs or implements it. For instance, *ambiguity* is the very nature of natural language (Piantadosi et al., 2012), and we have noted that even a low-agency, aligned task can have intrinsic disagreements as perceived by different people(Plank et al., 2014). This could be a gold mine for understanding human language, yet also add trouble to many machine learning setups with one "ground truth" label. Besides, many interesting cognitive patterns and biases per se – ones we might use collectively and unconsciously – are yet to be identified and/or understood. It would be insufficient to conclude that annotators are to blame as they apply such patterns or that researchers did not foresee these highly probable heuristics in their data. Included in the spectrum is an interesting case that confirms the Pollyanna hypothesis (our word choices include significantly more positive terms than negative ones) across languages (Dodds et al., 2015).