Whose Truth is it Anyway? An Experiment on Annotation Bias in Times of Factual Opinion Polarization

Abstract

Information shapes citizens' political decision-making. This process is amply studied by social scientists, for whom human annotation is a crucial instrument in their toolkit. Due to the democratization of data and the advances in NLP more data can be analyzed or classified, making these benchmarks more important than ever: An algorithm trained on biased data will reproduce and often exacerbate bias. Currently, disagreement and bias are often conflated—ignoring the possibility of annotator sample bias and valid disagreement. In a preregistered experiment in the Netherlands and a close replication in the U.S., we show that personal characteristics of annotators, like political ideology or knowledge, can interfere with annotators' judgement of political stances. Our results show that to improve annotated data for automated text analyses, and for stance detection models in particular, we need to critically evaluate how we create our gold standards.

Keywords: Experiment, Annotation Bias, Ideology, Measurinig Political Position, Text-as-Data, Political Knowledge

Introduction

Information shapes citizens' political decision-making. This relationship is extensively examined by social scientists at large, and political communication scholars in particular. Classical theories of political communication, such as agenda-setting and framing, highlight how political information influences opinion formation and participation in various political activities, from voting to protests (Van Aelst and Walgrave 2016; Lecheler and De Vreese 2019). The digital revolution has significantly expanded opportunities for creating and disseminating digital text, as well as for collecting and analyzing these sources. Recent overviews illustrate the wealth of new possibilities in this area (Van Atteveldt et al. 2022; Grimmer et al. 2022). Notably, advancements in Natural Language Processing (NLP) now enable the automatic analysis of vast amounts of data through machine learning techniques (Akyürek et al. 2020; Prost et al. 2019). However, this digital shift has also resulted in a complex media landscape characterized by high choice and fragmented audiences (Van Aelst et al. 2017; Fletcher and Nielsen 2017). Additionally, there is a growing trend toward polarization, with individuals increasingly interpreting information through the lens of their ideological beliefs (Rekker and Harteveld 2022; Lee et al. 2021).

We find ourselves in an era marked by an abundance of data and innovative analytical tools, yet simultaneously facing a decline in consensus and shared understanding regarding the events captured within that data. This has a significant impact on one of the core methods in political communication research: Text Analysis, whether manual or computational. This approach relies on the "consensual reading" of semantically or visually ambiguous messages by multiple coders across diverse contexts (Krippendorff 2004, p.212). In today's complex information landscape, it is no longer trivial for individuals to share the same contextual understanding when interpreting text. Consequently, the issue of measurement error—and how to address it—has garnered scholarly attention (Bachl and Scharkow 2017; Song et al. 2020; Aroyo and Welty 2015a). However, scholars make specific assumptions

here: a) There is a single correct interpretation of a text; b) Disagreement among annotators indicates unreliability, which is equated with invalid measurement; and c) Consensus among coders guarantees a valid interpretation of the instance in the real world, thus establishing a so-called "ground truth."

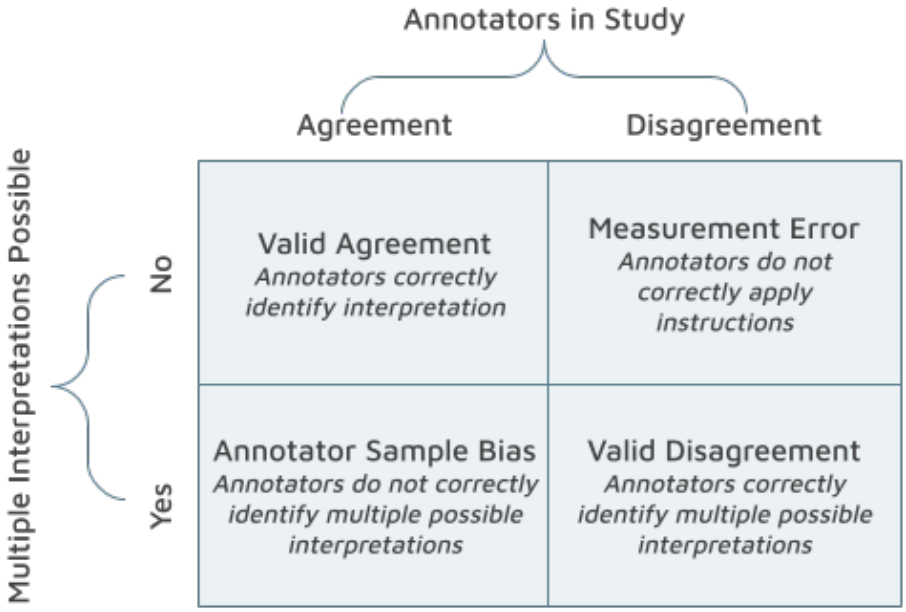
Based on recent literature, we content that these assumptions are no longer met. First, the NLP literature on perspectivism (Cabitza et al. 2023; Havens et al. 2022) and noisy labels (Zhan et al. 2019; Ibrahim et al. 2024; Khetan et al. 2017)—which involves incorporating the opinions and perspectives of human subjects during the knowledge representation phase of machine learning—suggests that people with differing ideological backgrounds or political leanings may interpret the same position in different ways (Shen and Rose 2021; Alkiek et al. 2022; Joseph et al. 2021). This challenges the idea of a single "ground truth" for every position and has sparked discussion on the practicality of assigning just one label to each example, especially when dealing with subjective or complex concepts and texts that allow for multiple interpretations. Interestingly, in experimental and survey-based studies on topics such as elite communication, researchers often account for heterogeneous treatment effects. This shows that we do not expect text-based treatments to have the same effect on different groups, such as partisans. Second, recent work (Baden et al. 2023) argues that some disagreement between coders should not be considered measurement error, but rather valid disagreement. Ambiguity in meaning or a lack of sufficient information—where annotators must rely on existing knowledge to "fill in the blanks"—means that disagreement does not necessarily indicate an incorrect measure, especially in current fragmented media landscape (Van Aelst et al. 2017; Fletcher and Nielsen 2017). Third, agreement among annotators does not always result in a valid gold standard (Song et al. 2020). Annotators from student populations or online platforms may not represent the general public (Clifford et al. 2015; Huff and Tingley 2015; Berinsky et al. 2014), potentially introducing bias if we insist on treating only one position as true.

Taken this together, we argue that based on level of consensus of the coders as well as on whether the text allows for multiple interpretations (textual ambiguity),

there are four possible outcomes–visualized in Figure 1–that need to be integrated when creating a gold standard for textual analysis methods. Figure 1 outlines four distinct outcomes that can arise during the annotation of political text, depending on two dimensions: (1) whether a shared ground truth is assumed or established, and (2) whether annotators agree with each other. To avoid ambiguity, we distinguish between two types of agreement: inter-annotator agreement (i.e., agreement among multiple annotators) and annotator–ground truth agreement (i.e., alignment between an annotator's label and a presumed correct label). The top-left quadrant represents cases where annotators agree with each other and with a presumed ground truth—what we refer to as valid agreement. This is the ideal scenario assumed in much of the traditional literature. The top-right quadrant represents cases where a ground truth is assumed but annotators disagree either with it or among themselves. In such cases, disagreement may indicate either measurement error (if annotators misunderstand the task or apply inconsistent logic) or bias (if systematic factors like ideology influence responses). In the bottom-left quadrant, annotators agree with each other, but their agreement may reflect a lack of diversity (e.g., if they share ideological backgrounds), leading to annotator sample bias. This is a case of high inter-annotator agreement but potentially low validity in terms of generalizability. Finally, the bottom-right quadrant captures valid disagreement, where multiple interpretations are justifiable due to inherent textual ambiguity. Here, annotators may diverge even when all are applying reasonable logic and effort, and thus, disagreement should not automatically be treated as error. This is described by Baden et al. 2023 as *valid disagreement*. Thus, Figure 1 presents a conceptual framework that outlines different types of disagreement that can arise during the annotation of politically charged text. Its purpose is to map the broader space of possible outcomes, distinguishing between forms of disagreement that result from systematic annotator bias, valid interpretive differences, or random error.

While not all quadrants of the figure are empirically tested in this study, it serves to position our contribution within this larger problem space. In our paper, we use a

Figure 1: Overview of Types of Biases.



population-based sample to address specific regions within the framework visualized in Figure 1. In addition, we provide guidance on how to deal with diversity in conceptions and political heterogeneity when creating gold standard data. We use a preregistered experiment in the Netherlands and a close replication in the U.S. (see here in the Netherlands—a low-level polarized country—and in the U.S.—a high-level polarized country.[1] Our first hypothesis (H1) focuses on systematic differences between annotators, such as ideological orientation and political knowledge, and how these may result in divergent interpretations—this relates to the quadrant capturing sample bias, where certain annotator groups interpret content differently in predictable ways. Our second hypothesis (H2) addresses the possibility of valid disagreement: when a sentence permits multiple reasonable interpretations, we test whether such divergence aligns with ideological background or knowledge level, in contrast to more unequivocal statements. This maps onto the part of the framework where multiple valid interpretations exist, rather than annotator error. Finally, our third hypothesis (H3) examines whether a design intervention—masking the

[1]Same procedure as pre-registered in the Netherlands is used for the U.S.

political actor—can reduce both systematic bias and valid disagreement, effectively targeting the central tension illustrated in the framework. Earlier work reports that coders of political texts incorporate prior beliefs about parties' issue stances into their coding decisions (Ennser-Jedenastik and Meyer 2018), concluding that party labels cue coders to a stance. For example, coders are more likely to report a left-wing party to be pro-immigration and a populist right-wing party to be against based on the exact same sentence. Additionally, we further examine the robustness of our findings by exploring: (a) the hypotheses for each issue separately, (b) interactions between the treatments, and (c) alternative measures of ideology.

Overall, our results demonstrate that for multi-interpretable sentences, political knowledge is applied. Respondents with higher levels of political knowledge were more likely to inferring a position where linguistically none is explicitly given. This demonstrates the need for a wider variety of annotators, since these sentences are very common in political text – like legislative debates or speeches – as well as media reports, and annotators with higher levels of political knowledge, like our students, pick on on the implicit position, but others do not. Moreover, our results also demonstrate that for these disagreements in the crowd to occur, the context is important, as we find differences in disagreements between the American and Dutch context. Our findings underline the importance of taking disagreement seriously for the creation of gold standard text – the bread-and-butter of all machine learning endeavors. We should look beyond the majority vote and model it in the data, because if an algorithm is trained on biased data from disagreeing annotators, it will reproduce and often exacerbate that bias when it is applied to new data (Prost et al. 2019, e.g.). We thus rather opt for modeling a diversity of viewpoints instead of choosing for (potentially) biased solving of disagreements. To be able to model these annotators' characteristics, we should survey the characteristics of annotators when using the crowd (see Webb-Williams et al. 2023 for a similar argument, yet different annotator characteristics).

Disagreement: Bias & Error

Generating large data sets has become one of the main drivers of progress in NLP (Akyürek et al. 2020; Piskorski et al. 2023). Studying political texts, the most familiar annotation tasks involve identifying theoretical concepts. Crowdcoding seems suitable for this task (Van Atteveldt et al. 2021). Yet, having only a few workers annotating the majority of texts has raised concerns about data diversity and models' ability to generalize beyond the crowd-workers (Clifford et al. 2015; Huff and Tingley 2015; Berinsky et al. 2014), since they tend to be younger and hold more liberal values than the general public. Often models do not generalize well to annotations from annotators that did not contribute to the training set (Geva et al. 2019). Moreover, a recent study in NLP suggest that experiential factors influence the consistency of how political ideologies are perceived: People with different ideological backgrounds seem to experience that position differently (Shen and Rose 2021). Especially in times of increasing political, affective, and opinion polarization (Boxell et al. 2022; Iyengar et al. 2019; Gidron et al. 2019; Rekker and Harteveld 2022), annotator sample bias (see Figure 1, based on a little variation in ideological position of the annotators, can occur.

Models automatically classifying texts are trained on data created by human annotators, including a final step where disagreements and differences in annotations are leveled by aggregating, averaging, or other ways of coming to a consensus on one label for one example—i.e. a "ground truth" (Aroyo and Welty 2015b). Differences from this ground truth label are seen as measurement error that need to be removed or sorted out. Models then learn to predict labels for new examples based on this ground truth. Recently, there is some discussion on valid disagreement in communication science (Baden et al. 2023) as well as on how realistic it is to have just one label for each example, especially for subjective or complex concepts and/or texts with multiple interpretations (Zhan et al. 2019; Ibrahim et al. 2024; Khetan et al. 2017).[2] Hence, disagreement can be informative for the researcher: It can for

[2]Disagreement even occurs in seemingly objective tasks such as Part of Speech tagging (Fornaciari et al. 2021; Plank et al. 2014).

instance be used to validate hypotheses about how universal the perceptions of such concepts are (Sommerauer et al. 2020). It can either indicate valid disagreement, suggesting that a diverse pool of annotators should be considered, or it may signal inconsistencies in annotation (measurement error) that require further clarification or refinement of coding guidelines.

Whose perspective is being recorded in these datasets, and then later in the models trained on these datasets? As we often use either students or crowd-coding (Van Atteveldt et al. 2021), we are faced with a lack of diversity in the annotator pool. This is undesirable in subjective and social tasks, such as hate speech detection or political affiliation classification. Framing arbitrary representations in data as "bias" misses the political character of data sets: There is no neutral data and no apolitical standpoint from which we can call out bias. Data sets are always "'a worldview' and, as such, data always remains biased" (Miceli et al. 2022, p.5). The answers of the annotators in turn influence how machine learning examples classify new models. For instance, hate speech and abuse detection are NLP tasks where race and gender of annotators influences both annotators and model performance (Gordon et al. 2022; Larimore et al. 2021; Waseem 2016).

Recent approaches in NLP have sought to explicitly incorporate disagreement and diversity in training data annotations, such as the explicitly subjective annotation paradigm (Röttger et al. 2022), "perspectivism" (Cabitza et al. 2023), "jury learning" (Gordon et al. 2022), and noisy labels (Zhan et al. 2019; Ibrahim et al. 2024; Khetan et al. 2017). In such approaches, demographic aspects of the annotator are explicitly mentioned and highlighted as having an influence on classification performance, but is used as an asset rather than as an error. A recent study suggest that extra-linguistic factors, such as annotators' political ideology and knowledge, potentially influence annotation and in turn model performance (Thorn Jakobsen et al. 2022). However, to our knowledge this phenomenon has not yet been tested in a controlled experimental setting with manipulations in the texts.

Some texts are ambiguous, and open to multiple interpretations. We call such sentences with more possible interpretations and less explicit standpoints "under-

specified". A lack of explicitness in the annotated text is one of the main causes of the disagreements in earlier literature: Annotator bias comes from a process known as the affect heuristic (Thorn Jakobsen et al. 2022), making a decision based on the emotional response related to your own personal attitude towards the discussed topic, especially when the text is relatively ambiguous (Slovic et al. 2007). We therefore we propose the following hypotheses:

Hypothesis 1a (Ideological Bias Hypothesis (H1a)): *The larger the ideological distance between respondent and the party, the less likely respondents annotate statements according to the party's uttered position.*

Hypothesis 1b (Ideological Bias Hypothesis (H1b)): *The effect of H1a is stronger for underspecified sentences.*

In the tradition of more strict interpretations of what constitutes as a stance especially in (computational) linguistics, sentences that are underspecified should always be annotated as "not a stance". Not only does this conflates measurement error and valid disagreement, it also ignores the idea of annotator sample bias. These become important notions when our main goal is to understand how readers are affected by statements shown in e.g., a newspaper article. In this case, we need a more lenient definition of stance that allows annotators to infer the direction of a political stance from context and prior knowledge even though a statement is strictly speaking underspecified might be more useful. In everyday life, people will not follow strict linguistic definitions, for understanding media effects we might thus be more interested in whether stances, giving context information, are "correctly overinterpreted". To test whether people with different ideological backgrounds as well as their political knowledge might experience underidentified position differently, challenging our ground-truth assumption in data annotation, we propose the following hypotheses:

Hypothesis 2a (Ideological Overinterpretation Hypothesis (H2a)): *The larger the ideological distance between respondent and the party, the more likely respondents interpret underspecified sentences as stance.*

Hypothesis 2b (Political Knowledge Overinterpretation Hypothesis (H2b)): *The more political knowledge, the more likely people interpret underspecified sentences as stance.*

Can we downplay both measurement error and valid disagreement, thereby increasing valid agreement? There have been some previous approaches to solve disagreement between annotators, especially where it concerns political or societal aspects. Geva et al. (2019) introduce an approach where training set annotators are separated from annotators annotating data sets that evaluate the models, to ensure the evaluation is not simply accurate at replicating the original annotators, but can generalize to new annotators' judgments. However, these approaches are not aimed at reducing biases during the training set annotation procedure. Other approaches are aimed at leveraging multiple perspectives - but this is not useful when one wants one label to learn from. For the Austrian National Elections, Ennser-Jedenastik and Meyer (2018) already demonstrated that showing party labels impact annotators' assessment of the party position. So, one solution is masking. We therefore test whether masking reduces potential differentiation incited by ideological position or political knowledge with the following hypotheses:

Hypothesis 3a (Masking Solution Hypothesis (H3a)): *Masking reduces the effect of respondents' ideological position for coding stances according to the party's position.*

Hypothesis 3b (Masking Solution Hypothesis (H3b)): *Masking reduces the effect of respondents' level of political knowledge for coding stances according to the party's position.*

Data, Methods & Measurement

We have conducted the survey experiments in the Netherlands in May 2022 and in the United States in January 2023 as we aim to test our hypotheses in both a

higher and lower polarized context to assess the broader applicability of how ideology and political knowledge influence the annotation of political text. The Dutch study is pre-registered. The U.S. experiment was designed as a close replication of the preregistered Dutch study. The procedure, including random assignment to experimental conditions, survey flow, outcome measures, and analysis strategy, is identical to that used in the Netherlands. The only difference lies in the wording of the statements, which were adapted to reflect the U.S. political context and party system. These adaptations ensured linguistic and political relevance while preserving the structure and ambiguity levels of the original Dutch items. The full set of statements used in both experiments is presented in Table 1. Both samples, recruited through *KiesKompas* and Prolific for respectively the Dutch and American case, consist of 2,000 participants (based on the power analysis presented in our online compendium,[3]. Both survey companies works with non-random opt-in respondents. Therefore, we measured many demographic background variables, and balance checks have been conducted to demonstrate whether certain categories are over represented in a certain experimental group (see OA A-3). Our study has been approved by the Research Ethics Review Committee of the researchers' institution (see here. To ensure good quality of our data, one attention check (discussed in more detail in OA A-2) is included (Berinsky et al. 2014).

Measurement

Experimental Conditions. Respondents are randomly assigned to either view a political party as an actor, or a masked condition, where they see 'X' as an actor; simultaneously, respondents see either a fully specified sentence or an underspecified sentence, in which one needs additional information outside of the text to determine an actor's position. Table 1 gives an overview of the variations in treatment in the surveys. All specified sentences come from the Dutch and American Electoral Compasses, the underspecified sentences are adapted. Due to the nature of the different party systems, the sentences in the Dutch context uses more political

[3]Same procedure is used for the U.S.

parties.

Table 1: Survey Questions - Experimental Conditions.

Condition	**US Experiment**	**NL Experiment**
Specified	[Republicans/X] say immigration should be made more difficult.	[PVV/X] says immigration should be made harder.
Specified	[Democrats/X] say we need to put a tax on carbon emissions.	[GreenLeft/X] says nitrogen emissions need to be reduced.
Specified	[Democrats/X] say we should implement a wealth tax for the richest Americans.	[Labour Party/X] says tax rate should go up for highest earners.
Specified	[Republicans/X] say the U.S. needs to consider military build-up in the Pacific Ocean.	[Forum for Democracy/X] says that membership in the European Union has been especially bad for the Netherlands so far.
Underspecified	[Republicans/ X] say many immigrants are crossing our borders.	[PVV/X] says many immigrants are coming this way.
Underspecified	[Democrats/X] say carbon emissions policy should be implemented differently.	[GreenLeft/X] says nitrogen policy must be different.
Underspecified	[Democrats/X] say the tax system should be implemented differently.	[Labour Party/X] says tax system must be changed.
Underspecified	[Republicans/ X] say there should be a different military presence in the Pacific Ocean.	[Forum for Democracy/X] says the Netherlands should have a different role in the European Union.

Dependent Variable. We rely on four dependent variables to assess how respondents interpret a party's (implied) stance: (1) *Correct Stance Identification - One Perspective* (Strict Interpretation): Whether respondents correctly identify the party's position based on a strict definition of stance, which requires both a specification of change and direction. Respondents are considered correct if they answer "no stance" for underspecified sentences and "against," "in favor," "in favor," and "against" for specified sentences one through four; (2) *Correct Stance Identification - Multiple Perspective* (Lenient Interpretation): Whether respondents correctly identify the party's position based on a broader definition of stance, allowing for either "in favor" or "against" responses for underspecified sentences two to four, as long as they recognize a directional stance; (3) *Over-interpretation of Stance*:

Whether respondents incorrectly infer a stance in cases where none is explicitly stated, which serves as an indicator of measurement error; and (4) *Stance Recognition Failure*: Whether respondents fail to identify a stance when one is explicitly provided, indicating potential annotator sample bias. For each issue, we ask the respondent *what is according to the sentence above the position of [ACTOR]?*', with the answer categories: `in favor`, `against`, `no stance`, `don't know`. These variables allow us to examine the extent to which disagreement and ambiguity affect stance interpretation across different levels of specification. It is important to note that in the pre-registration, we specified only the strict identification of stances. However, in the results, we also applied a more lenient interpretation. This constitutes a deviation from the original pre-registered analysis plan.

Moderating Covariates. *Ideological position* is measured using an 11-point scale ranging from left (0) to right (10), as asked in the survey. For the analyses, we transform this into a measure of ideological distance, defined as the absolute difference between the respondent's self-placement and the position of the political party featured in the sentence. Party positions are derived from the Electoral Compass, based on expert-coded stances for each statement. This transformation captures how ideologically close or distant a respondent is from the party making the statement, which is central to our theoretical expectations. *Political knowledge* is measured with the standard six items from the Dutch Parliamentary Election Studies for the Dutch sample, and the standard three items from the American National Election Studies.[4]

Control Variables. As pre-registered, we have conducted a balance test (see Appendix A-3). This showed that none of our coviarates where unbalanced over the treatment groups. We therefore have not included any control variables.

Method

To test our hypotheses, we will conduct a multilevel model, with respondents clustered in sentences, where issues (topics) are entered as fixed effects, consistent with

[4]The questionnaire can be found in OA A-1.

the preregistration The standard errors are clustered at the individual level. Given the binary nature of several outcome variables (e.g., correct stance identification, overinterpretation), we model these using linear probability models (LPMs). We chose LPMs over logistic regression for reasons of interpretability: the coefficients from LPMs are directly interpretable as changes in predicted probability, which we believe enhances clarity for our audience. While we acknowledge that logistic models offer advantages (e.g., bounded predictions), we have verified that our key findings are robust to using logistic regression. For transparency, we note that our results and interpretations do not hinge on the specific modeling framework chosen.

Deviations from Preregistration

First, in the preregistration, only a strict interpretation of stance was planned—requiring both direction and a change to be present in the text. In the current manuscript, we additionally report results based on a more lenient interpretation of stance, which allows for directional inferences even when the sentence lacks a clear policy change. This addition reflects the observation that many real-world political statements are under-specified, and thus captures meaningful variation in how annotators interpret such content.

Second, while the U.S. experiment was not preregistered, it closely replicates the Dutch design. We analyze both studies using the same statistical models, but clearly identify the U.S. study as a post hoc extension of the preregistered work.

Third, we introduced several exploratory analyses not included in the preregistration. These include interaction effects (e.g., between masking and sentence type), issue-specific subgroup analyses, and alternative operationalizations of ideological distance and political knowledge. These are labeled as exploratory throughout the manuscript.

Fourth, the preregistered power calculation was based on a simplified OLS regression model applied to each sentence separately. In the final analysis, we opted for multilevel models pooling across issues and clustering standard errors at the

respondent level. While the sample size remains adequate, we acknowledge that the preregistered power analysis does not fully reflect the final analytical approach.

Results

Before discussing the results of our preregistered analysis, we will describe which profiles, based on ideology (see Figure 2) and political knowledge (see Figure 3), are more likely to exhibit bias or introduce multiple perspectives. OA A-5 gives even more detail; it describes the average profile of the respondents who agree and disagree during annotations in percentage. Figure 2 demonstrates that in both countries people positioned on the ideological left of the spectrum are more likely to agree in specified sentences with our gold standard (in green) – i.e. testing the upper-right (valid agreement) and upper-left (measurement error) of Figure 1. In general, we see that for the specified sentences, the coders where almost always in line with our gold standard. This picture is a bit different for the underspecified sentences. In principle, all these sentences where defined so that they are linguistically neutral (no stance, in blue). However, in all cases except for the sentence 'many immigrants are crossing the border' – when the party is masked – a stance is still interpreted. Except for the sentences "there should be a different military presence in the Pacific Ocean" and "carbon emissions should be implemented differently" when the Democrats are shown, all sentences are interpreted as negative. Interestingly, in underspecified sentences—similar to those commonly found in political communication studies—there is significant variation in how they are coded, depending on the annotator's ideological position. Looking at the profiles based on political knowledge, Figure 3 demonstrates that in the US, people with lower levels of political knowledge are more likely to be correct, whereas in the Netherlands, people with higher levels of political knowledge are more likely to be correct. This indicates that biases based on political knowledge can be context-dependent. For the rest, Figure 3 paints a similar picture as 2: Respondents where to high degrees correct based on the gold standard in the specified sentences, but interpreted the

underspecified sentences overwhelmingly as a negative stance.

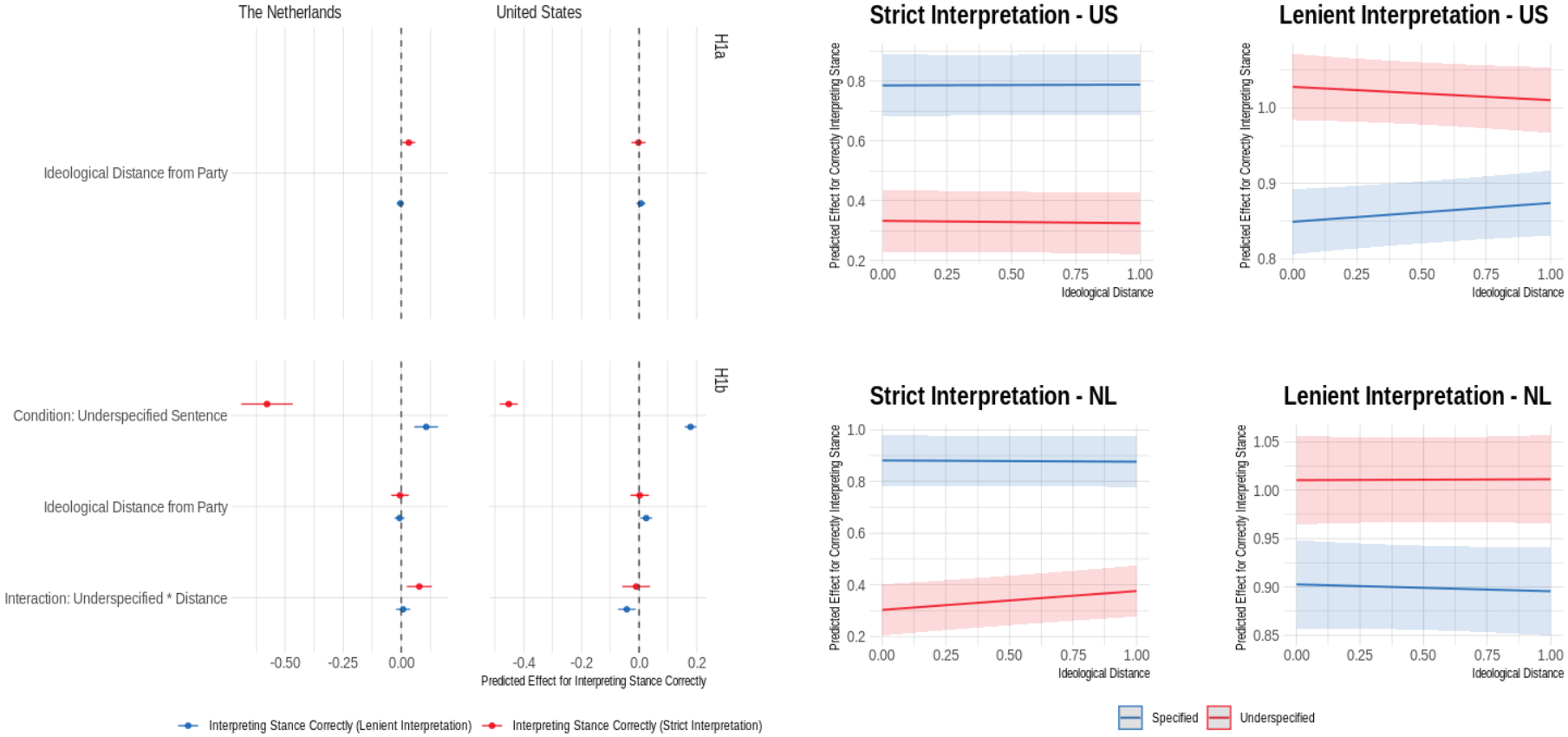
Figure 2: Profile Ideological Distance.

To answer whether there is ideological or knowledge-based annotation bias, we have conducted a two-by-two experiment. To test our hypotheses, we will conduct a multilevel model, with respondents clustered in issues (for more details, see OA A-4). OA A-6 displays the effect of the experimental conditions on the four dependent variables – correctly identifying a stance and over-interpreting a stance for both a strict and a lenient interpretation of stance. OA A-7 demonstrates the exploratory analyses that test the robustness of our results presented below.

Figure 3: Profile Political Knowledge.

Figure 4: Regression results of overinterpreting stance, by ideological distance and sentence specificity.

((a)) Regression Coefficients. **((b))** Predicted Effects.

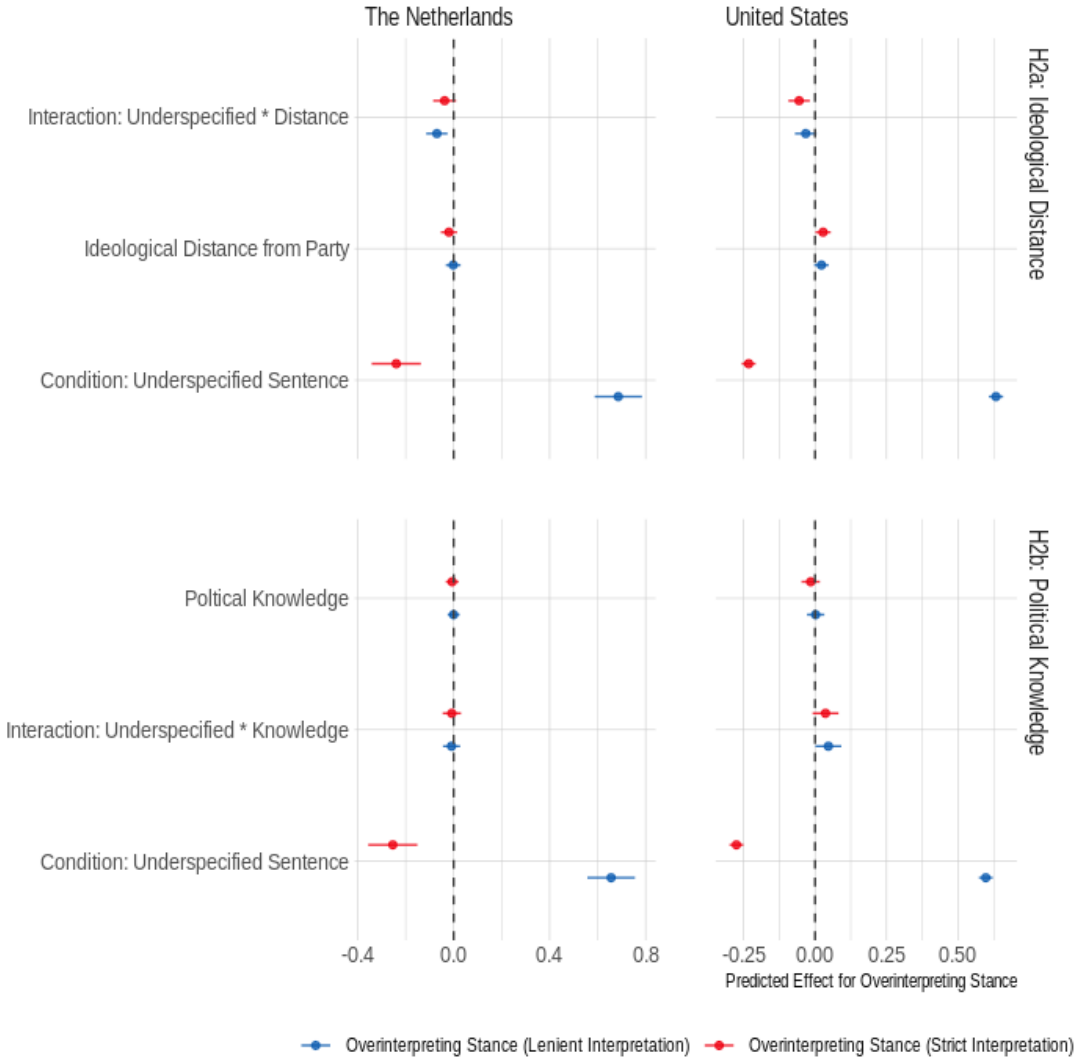


Notes: Coefficients are derived from a multilevel linear probability model with respondents nested in sentences. The dependent variable is overinterpretation of stance (coded as 1 if a stance is inferred from an underspecified sentence, 0 otherwise). Error bars represent 95% confidence intervals. Predictions are computed while holding other covariates—political knowledge and masking condition—at their sample means (for continuous variables) or reference categories (for categorical variables).[5]..

First, we test whether there is an ideological bias in interpreting stances (H1a), and if this bias increases for those that are further away from the ideological position of the political actor in the under-specified condition (H1b). Figure 4 shows the regression results of overinterpreting a stance, based on a respondent's ideological distance from the party and whether the sentence is specified or underspecified. The goal is to assess whether people further from the party are more likely to infer a stance even when the sentence is ambiguous. This plot allows us to compare effects across sentence types. Figure 4(a) demonstrates on the regression coefficients and Figure 4(b) the predicted effects. The upper-left panel of Figure 4(a) demonstrates the coefficient of ideological distances for the likelihood of interpreting the stance correctly. There is a negligible positive effect – a coefficient of '0.002' – that is borderline significant in the Netherlands, when looking at the strict interpretation of a stance. Substantially, this means that there is no effect of ideological distance for correctly interpreting the stance in either the American or Dutch case. Hence, no ideological bias found, thus no support for our H1a. Looking at the lower-left panel of Figure 4(b), we see a small but significant effect of the interaction between ideological distance and the under-specified condition for the strict interpretations in the Dutch case (hypothesized direction), and a small negative effect for the lenient interpretation of a stance in the US case (other direction as hypothesized). Those who are ideologically further away from the party are less likely to be wrong than those who are close to the party in the Netherlands, as shown in the right-hand panel of Figure 4(b). For the Dutch case, using the strict interpretation, the difference is about `10%` – from a coefficient of `0.3` to `0.4`. In the American case, the slope for the lenient interpretation goes down as predicted, but not that much, with a difference of about `5%`. This could indicate that in times of heightened polarization, ideological annotation bias can be a concern. We thus find mixed evidence for H1b.

Secondly, we hypothesized that there is a risk of over-interpretation from those that are ideologically distant to the political actor (H2a) as well as those who have high levels of political knowledge (H2b). We measure this with an interaction be-

Figure 5: Predicted effects of overinterpreting political stance, by ideological distance and political knowledge.



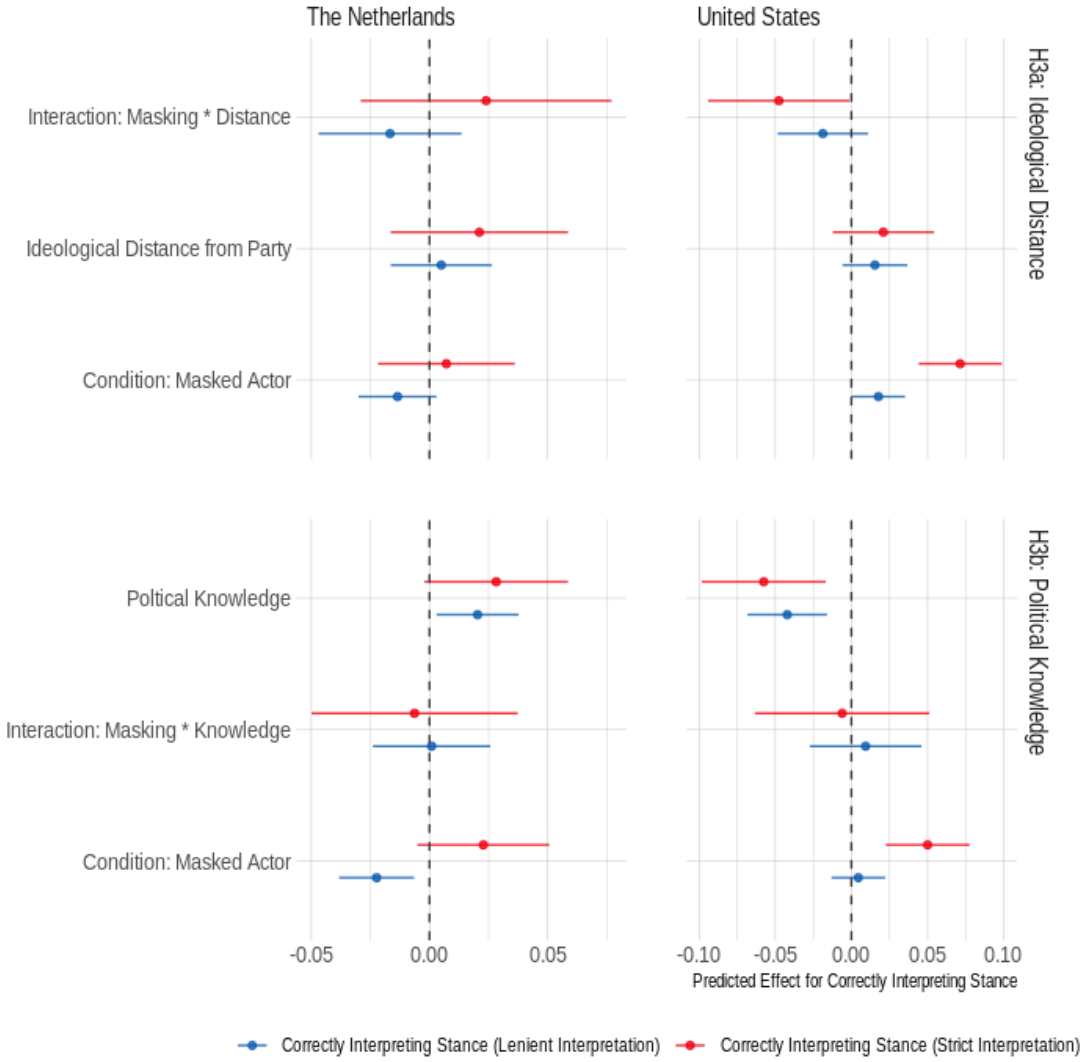
Notes: Figure displays predicted probabilities of overinterpreting stance (i.e., inferring a stance when the sentence is underspecified) based on interaction effects between ideological distance (left panel) and political knowledge (right panel) with sentence specificity. Predictions are based on multilevel linear probability models with respondents nested within sentences. Error bars indicate 95% confidence intervals. Other covariates are held at their sample means or reference categories when generating predictions.[6]..

tween the experimental condition of specification and the variable of interest. To assess whether ideological distance or political knowledge contributes to overinterpretation of stance in ambiguous sentences, Figure 5 plots predicted probabilities

of overinterpreting stances across these two variables. The left panel focuses on ideological distance, while the right panel shows the role of political knowledge. In both cases, we compare results for specified versus underspecified sentences. This allows us to test H2a and H2b across both national contexts. In both countries, we see that the different interpretations of stance have opposite effects: A strict interpretation decreases the chance of overinterpreting, and this is exacerbated by ideological distance but not by political knowledge. Using a more lenient interpretation, we see that this increases the chance of overinterpreting. While this is not further diminished by political knowledge, ideological distance does further diminish the change in the American case – i.e. against expectation H2b. In the Dutch case, however, the further you are from the party the more likely you are to overinterpret the sentence as a stance. So, while we do find some support for our H2a, we will reflect on the interpretation of a stance in combination with the context, as this effects how crowds interpret underspecified sentences.

Thirdly, we test whether masking of the political actor is a solution for potentially misinterpreting the stance. We hypothesized that masking should reduce the ideological bias (H3a) as well as the bias resulting from political knowledge (H3b). Figure 6 tests our third hypothesis (H3), which posits that masking the political actor may reduce ideological or knowledge-based bias in stance interpretation. The top panels examine whether masking reduces the effect of ideological distance on stance coding, while the bottom panels assess its impact on differences driven by political knowledge. Each panel shows regression estimates with confidence intervals for the interaction effects in both the Netherlands and U.S. contexts. There is no effect found in the American or Dutch case regarding ideological distance. The same goes for the interaction between masking and political knowledge, on the bottom-panel of Figure 6. There is no significant effect found in the US nor in the Netherlands when we look at the strict interpretation. Looking at the lenient information, allowing for valid disagreement, we do find that not masking has a significant negative effect for the Dutch case for the variable political knowledge. This means that masking of political actors does not help to correctly interpret a

Figure 6: Effect of masking political actors on stance interpretation: interactions with ideology and political knowledge.



Notes: Figure displays regression coefficients and 95% confidence intervals from multilevel linear probability models estimating the effect of masking the political actor on correct stance identification and overinterpretation. Separate panels show interactions between the masking condition and (1) ideological distance and (2) political knowledge. Error bars indicate 95% confidence intervals. Other covariates are held at their sample means or reference categories when generating predictions.[7]..

sentence as a stance if one assumes a ground truth – i.e. no support for H3a and H3b for the strict interpretation – but some support for H3b if one acknoledges valid disagreement.

Exploratory Results

To check the robustness of our findings, Figure OA.4 demonstrates the analyses for each issue separately. None of these analyses are pre-registered. The different colors visualize the different dependent variables. We do not see much variation between issues *Tax*, *EU/Foreign Policy*, and *Environment*. For those issues, we see that almost everyone interprets the sentence correctly (in blue and red). We also see that for a lenient interpretation of stances, people are quite likely to overinterpret a position as a stance. Being correct about the stance does not decrease when masking the political actor in both cases. Yet, the chance of being correct decreases statistically significantly when the sentence is underspecified. The same holds for overinterpreting for the lenient interpretation, but the opposite is true for the strict interpretation; there overinterpretation is more likely with underspecified sentences. Looking at *Immigration*, we see a different pattern. We see that masking does not increase the likelihood of being correct, but does increase the likelihood of overinterpretation regardless of how one defines a stance. Underspecified sentences are less likely to be correctly identified and more likely to be overinterpreted regardless of the definition of a stance. So, while there are some differences in effect sizes between the issues, the overall findings are not driven by a single issue.

In addition to issue-specific analyses, we also explore an interaction between treatments, visualized in Figure OA.5 for both dependent variables. This shows that masking is of help when sentences are under-specified. In the left-hand panel of Figure OA.5, it demonstrates that for under-specified sentences, people are less likely to incorrectly identify a sentence as a stance when the actor is masked (coefficient of `-0.30`) than when an actor is revealed (coefficient of `-0.45`). That means there is a 15% increase in having it correct. The difference for overinterpreting is smaller between revealed and masked political actors – shown in the right-hand panel of Figure OA.5 – yet also statistically significant. Compared to `85%` overinterpreting the sentence as a stance, in the masking solution "only" `80%` over-interprets the sentence as a stance. In the recommendation section, we will reflect on the

masking solution for under-specified sentences.

Lastly, we explore three different ways of measuring ideological distance and an alternative for political knowledge in the American case. First, we measured ideological bias by looking at whether the respondent is congruent or not with the issue position in the sentence, visualized in Figure OA.7. Second, we measured ideological bias by looking at whether the person voted for the party displayed in the sentence, visualized in Figure OA.6. Thirdly, we measured ideological bias by looking at the ideology of the respondents – not in relation to the political actor revealed, visualized in Figure OA.6. These figures show that our null-finding regarding ideological bias is not conditional upon the measure we used. In none of the analyses, we find evidence for ideological bias. Also for the alternative measurement of political knowledge in the US, we find the same results as reported in the main analyses.

Conclusion

Human annotation represents a crucial instrument in the toolkit of the social sciences. In order to address concerns regarding the validity and reliability of annotation tasks, we establish strict standards. Especially in times where due to the democratization of data and the advances in NLP more data can be analyzed or classified, these standards are more important than ever. After all, an algorithm trained on biased data will reproduce and often exacerbate bias (Prost et al. 2019, e.g.). It is key to realize that the dominant underlying assumption to create valid and reliable annotated data is that there is only one correct interpretation for every input example, and that disagreement between the annotators is something that needs to be dealt with at all costs (Aroyo and Welty 2015b). In times of increasing factual opinion polarization (Rekker and Harteveld 2022; Lee et al. 2021), it is crucial to ask *whose perspective is being recorded in these datasets*, given that there is no neutral data and no apolitical standpoint from which we can call out bias. Recent debates in NLP challenge the assumption that disagreements among annotators are

purely errors to be corrected. Instead, these 'noisy labels' can reflect genuine diversity of perspectives, especially in subjective or complex annotation tasks (Zhan et al. 2019; Ibrahim et al. 2024; Khetan et al. 2017). Studies suggest leveraging such disagreements as informative signals, integrating multiple viewpoints to enhance model robustness and fairness rather than enforcing a singular 'ground truth' (**cabitza2023perspectivism**; Baden et al. 2023). Our tools to create valid and reliable annotation data does currently not allow for dealing with different perspectives. In this paper, we examined annotation bias in the classification of political stances using a preregistered experiment in the Netherlands and a close replication in the U.S. We used both a strict and a lenient interpretation of stance to illustrate whether decisions made in the research process on what constitutes a stance influences the results.

First, we tested whether ideological bias affects the interpretation of a party's stance by analyzing whether a person's ideological distance to a party affects how they classify the party's stance. Although we did not find a main effect of ideological distance (H1a), we did find evidence that the effect of ideological distance is greater for underspecified sentences (H1b) in the US, depending on how one operationalizes stances. The exploratory analyses (see OA A-7) showed that these findings where robust against different specifications of ideological distance. This finding suggests potential annotator sample bias in highly polarized contexts, where ideological divisions may lead coders to systematically perceive and interpret the same position differently, rather than recognizing a single or multiple perspectives as intended. This conditional finding is important, because machine translation is on the rise (Licht 2023; De Vries et al. 2018): Social scientists use this technique for example to comparatively analyze political stances. Currently, most scholars using text analysis rely on English text or annotated data, translating the original source text to English (Baden et al. 2022). Our results show that the way a political stance is defined and the context, potentially the level of polarization, affects whether annotators induced bias. Hence, naively using US annotated data and auto-translating your source text could have downstream consequences for the

classification of your stances.

Second, we looked at differences in terms of overinterpretation, which we define as coding stance in an underspecified sentence (i.e. in which the stance is not explicitly made clear) using both a strict (computational linguistics) and more lenient interpretation of what a stance is. In other words, we tested when annotators inferred the stance of the party on the issue based on prior knowledge or beliefs about the party. We expected that overinterpretation is more likely if the ideological distance of the coder to the party is greater (H2a) and if the coder has more political knowledge (H2b). Overall, our findings did not support these hypotheses. Yet, they imply that ideological distance can affect how stance is processed. With the more lenient interpretation of overinterpretation, we did find support for H2a for the Dutch context, but we found an opposite effect for the US. While this does not support the hypothesis in terms of the direction of the bias, it does imply that ideological distance can affect how stance is processed. Regarding political knowledge (H2b), our results showed again country differences as well as that the way one defines a stance matters. Our findings, for H2b in particular, pose important implications. For sentences with clear political stances on an issue (e.g., immigration should be made more difficult) annotators generally agreed on the stance that is expressed. But if the sentence refers to the issue without making a clear stance (e.g., many immigrants are crossing our borders) we do see a clear effect of political knowledge. This indicates that for annotation bias to occur, there has to be sufficient room for different interpretations. Hence, the concept of valid disagreement (Baden et al. 2023) should be integrated into our practices.

If the stance is clearly specified, then there is more of a ground truth, and the opinions and perspectives of the annotator therefore matter less. It is when the stance is underspecified that perspectivism (Cabitza et al. 2023; Havens et al. 2022) comes into play. When a party talks about a political issue without taking an explicit stance, annotators need to fill in the gaps and make their own inferences about the party's stance on this issue. Considering that in many political texts, specifically reporting on political news, stances are not made explicit, underspecified

sentences can be assumed to rather be the rule than the exception. Our results therefore imply that political knowledge is indeed a relevant dimension in shaping how an annotator perceives political stances. For this, either a Bayesian framework or active learning could be potential solutions. Given the uncertainty inherent in political stance annotation, a Bayesian approach could explicitly models this uncertainty in the learning process. Alternatively, active learning techniques could be used to identify the most informative or controversial examples for annotation.

Finally, we investigated whether either measurement error or valid disgreement due to ideological distance from the party (H3a) and political knowledge (H3b) could be alleviated by masking the political party. Here we again found country differences as well as variation based on how a stance was defined. We showed that masking has no effect to resolve disagreement in the US context, but it does help in a less polarized context like the Netherlands when you allow for valid disagreement between coders. Given that the US is one of the front runners in terms of polarization, it is important to realize that this affects our annotators, and therefore annotated data, too. Disagreement seem to be stronger and less easy to resolve.

Our results imply that to improve annotated data for automated text analyses, and for stance detection models in particular, we need to critically evaluate how we create our gold standards. In the best case scenario, ignoring personal differences on dimensions such as political ideology only adds noise to the data. However, this is only the case if a sufficiently large sample of annotators is randomly drawn from the population. In reality, gold standards are often the product of a handful of expert coders, who typically study or work at universities. Even if a good amount of crowd coders is used, it cannot blindly be assumed that this provides a diverse and balanced reflection of political ideologies in society, as we also observed in our samples. Especially in times where an increasing amount of crowd coding answers are based on large language models such as ChatGPT which (by and large) simulate the same ideological position over and over again (Veselovsky et al. 2023), reflecting on the composition and quality of annotators and how it influences gold standards becomes more than just a methodological exercise. Our results indicate

that an annotation team that is skewed towards one end of the political spectrum could indeed result in a biased gold standard due to either not recognizing valid disagreement, i.e. suffering from annotator sample bias, or by not modeling valid disagreement.

Furthermore, if the perception of political stance is contingent on ideological differences, then it can be inherently problematic to conceptualize political stance, as expressed in a text, as a ground truth. Depending on the goal of a study, it might be more appropriate to conceptualize political stance as having an inherently subjective quality, and strive to measure this subjective space. As a general guideline, we propose making a distinction between content analysis research that makes inferences about the author of a text, and research that makes inferences about the audience. For example, a study that compares political stances between news outlets to analyze political bias in the news does require a consistent and reliable measurement of political stances. This could indeed require a ground-truth definition of political stance, that is drilled into annotators through well-defined and specific instructions and extensive training. But if the study involves making inferences about how the audience interprets political stances, for example in media effects research, then a ground-truth definition of political stance is problematic. To make accurate inferences about how a citizen subjectively perceived political content, and consequently how this might have affected behavior such as voting, we need to learn more about how shared personal characteristics relate to shared modes of perspective.

References

Akyürek, Afra Feyza et al. (July 2020). "Multi-Label and Multilingual News Framing Analysis." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8614–8624. URL: https://aclanthology.org/2020.acl-main.763.

Alkiek, Kenan, Bohan Zhang, and David Jurgens (2022). "Classification without (Proper) Representation: Political Heterogeneity in Social Media and Its Implications for Classification and Behavioral Analysis." In: *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 504–522.

Aroyo, Lora and Chris Welty (2015a). "Truth is a lie: Crowd truth and the seven myths of human annotation." In: *AI Magazine* 36.1, pp. 15–24.

— (Mar. 25, 2015b). "Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation." In: *AI Magazine* 36.1. Number: 1, pp. 15–24. URL: https://ojs.aaai.org/index.php/aimagazine/article/view/2564.

Bachl, Marko and Michael Scharkow (2017). "Correcting measurement error in content analysis." In: *Communication Methods and Measures* 11.2, pp. 87–104.

Baden, Christian et al. (2022). "Three gaps in computational text analysis methods for social sciences: A research agenda." In: *Communication Methods and Measures* 16.1, pp. 1–18.

Baden, Christian et al. (2023). "Meaning multiplicity and valid disagreement in textual measurement: A plea for a revised notion of reliability." In: *SCM Studies in Communication and Media* 12.4, pp. 305–326.

Berinsky, Adam J, Michele F Margolis, and Michael W Sances (2014). "Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys." In: *American Journal of Political Science* 58.3, pp. 739–753.

Boxell, Levi, Matthew Gentzkow, and Jesse M Shapiro (2022). "Cross-country trends in affective polarization." In: *Review of Economics and Statistics*, pp. 1–60.

Cabitza, Federico, 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*. Vol. 37. 6, pp. 6860–6868.

Clifford, Scott, Ryan M Jewell, and Philip D Waggoner (2015). "Are samples drawn from Mechanical Turk valid for research on political ideology?" In: *Research & Politics* 2.4, p. 2053168015622072.

De Vries, Erik, Martijn Schoonvelde, and Gijs Schumacher (2018). "No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications." In: *Political Analysis* 26.4, pp. 417–430.

Ennser-Jedenastik, Laurenz and Thomas M Meyer (2018). "The impact of party cues on manual coding of political texts." In: *Political Science Research and Methods* 6.3, pp. 625–633.

Fletcher, Richard and Rasmus Kleis Nielsen (2017). "Are news audiences increasingly fragmented? A cross-national comparative analysis of cross-platform news audience fragmentation and duplication." In: *Journal of communication* 67.4, pp. 476–498.

Fornaciari, Tommaso et al. (June 2021). "NAACL-HLT 2021." In: Online: Association for Computational Linguistics, 2591–2597. URL: https://aclanthology.org/2021.naacl-main.204.

Geva, Mor, Yoav Goldberg, and Jonathan Berant (2019). "Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Gidron, Noam, James Adams, and Will Horne (2019). "Toward a comparative research agenda on affective polarization in mass publics." In: *APSA Comparative Politics Newsletter* 29, pp. 30–36.

Gordon, Mitchell L et al. (2022). "Jury learning: Integrating dissenting voices into machine learning models." In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–19.

Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Havens, Lucy et al. (2022). "Beyond Explanation: A Case for Exploratory Text Visualizations of Non-Aggregated, Annotated Datasets." In: *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pp. 73–82.

Huff, Connor and Dustin Tingley (2015). ""Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents." In: *Research & Politics* 2.3, p. 2053168015604648.

Ibrahim, Shahana et al. (2024). "Learning From Crowdsourced Noisy Labels: A Signal Processing Perspective." In: *arXiv preprint arXiv:2407.06902*.

Iyengar, Shanto et al. (2019). "The origins and consequences of affective polarization in the United States." In: *Annual review of political science* 22, pp. 129–146.

Joseph, Kenneth et al. (2021). "(Mis) alignment Between Stance Expressed in Social Media Data and Public Opinion Surveys." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 312–324.

Khetan, Ashish, Zachary C Lipton, and Anima Anandkumar (2017). "Learning from noisy singly-labeled data." In: *arXiv preprint arXiv:1712.04577*.

Krippendorff, Klaus (2004). "Reliability in content analysis: Some common misconceptions and recommendations." In: *Human communication research* 30.3, pp. 411–433.

Larimore, Savannah et al. (2021). "Reconsidering Annotator Disagreement about Racist Language: Noise or Signal?" In: *Proceedings of the Ninth*

International Workshop on Natural Language Processing for Social Media, pp. 81–90.

Lecheler, Sophie and Claes H De Vreese (2019). *News framing effects: Theory and practice*. Taylor & Francis.

Lee, Nathan et al. (2021). "More accurate, but no less polarized: Comparing the factual beliefs of government officials and the public." In: *British Journal of Political Science* 51.3, pp. 1315–1322.

Licht, Hauke (2023). "Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings." In: *Political Analysis*, pp. 1–14.

Miceli, Milagros, Julian Posada, and Tianling Yang (2022). "Studying up machine learning data: Why talk about bias when we mean power?" In: *Proceedings of the ACM on Human-Computer Interaction* 6.GROUP, pp. 1–14.

Piskorski, Jakub et al. (2023). "Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup." In: *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pp. 2343–2361.

Plank, Barbara, Dirk Hovy, and Anders Søgaard (2014). "Learning part-of-speech taggers with inter-annotator agreement loss." In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 742–751.

Prost, Flavien, Nithum Thain, and Tolga Bolukbasi (2019). "Debiasing Embeddings for Reduced Gender Bias in Text Classification." In: *GeBNLP 2019* 9573, p. 69.

Rekker, Roderik and Eelco Harteveld (2022). "Understanding factual belief polarization: the role of trust, political sophistication, and affective polarization." In: *Acta Politica*, pp. 1–28.

Röttger, Paul et al. (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*, pp. 175–190.

Shen, Qinlan and Carolyn Rose (2021). "What sounds "right" to me? experiential factors in the perception of political ideology." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1762–1771.

Slovic, Paul et al. (2007). "The affect heuristic." In: *European journal of operational research* 177.3, pp. 1333–1352.

Sommerauer, Pia, Antske Fokkens, and Piek Vossen (2020). "Would you describe a leopard as yellow? evaluating crowd-annotations with justified and informative disagreement." In: *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4798–4809.

Song, Hyunjin et al. (2020). "In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis." In: *Political Communication* 37.4, pp. 550–572.

Thorn Jakobsen, Terne Sasha et al. (June 2022). "LAW-LREC 2022." In: Marseille, France: European Language Resources Association, 44–61. URL: https://aclanthology.org/2022.law-1.6.

Van Aelst, Peter and Stefaan Walgrave (2016). "Political agenda setting by the mass media: Ten years of research, 2005–2015." In: *Handbook of public policy agenda setting*, pp. 157–179.

Van Aelst, Peter et al. (2017). "Political communication in a high-choice media environment: a challenge for democracy?" In: *Annals of the International Communication Association* 41.1, pp. 3–27.

Van Atteveldt, Wouter, Damian Trilling, and Carlos Arcila Calderon (2022). *Computational Analysis of Communication*. John Wiley & Sons.

Van Atteveldt, Wouter, Mariken ACG Van der Velden, and Mark Boukes (2021). "The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms." In: *Communication Methods and Measures* 15.2, pp. 121–140.

Veselovsky, Veniamin, Manoel Horta Ribeiro, and Robert West (2023). *Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks*. arXiv: 2306.07899 [cs.CL].

Waseem, Zeerak (Nov. 2016). "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter." In: Austin, Texas: Association for Computational Linguistics, 138–142. URL: https://aclanthology.org/W16-5618.

Zhan, Xueying et al. (2019). "Learning from multi-annotator data: A noise-aware classification framework." In: *ACM Transactions on Information Systems (TOIS)* 37.2, pp. 1–28.