Judgment Sieve: Reducing Uncertainty in Group Judgments through Interventions Targeting Ambiguity versus Disagreement

QUAN ZE CHEN, University of Washington, USA
AMY X. ZHANG, University of Washington, USA

When groups of people are tasked with making a judgment, the issue of uncertainty often arises. Existing methods to reduce uncertainty typically focus on iteratively improving specificity in the overall task instruction. However, uncertainty can arise from multiple sources, such as ambiguity of the item being judged due to limited context, or disagreements among the participants due to different perspectives and an under-specified task. A one-size-fits-all intervention may be ineffective if it is not targeted to the right source of uncertainty. In this paper we introduce a new workflow, Judgment Sieve, to reduce uncertainty in tasks involving group judgment in a targeted manner. By utilizing measurements that separate different sources of uncertainty during an initial round of judgment elicitation, we can then select a targeted intervention adding context or deliberation to most effectively reduce uncertainty on each item being judged. We test our approach on two tasks: rating word pair similarity and toxicity of online comments, showing that targeted interventions reduced uncertainty for the most uncertain cases. In the top 10% of cases, we saw an ambiguity reduction of 21.4% and 25.7%, and a disagreement reduction of 22.2% and 11.2% for the two tasks respectively. We also found through a simulation that our targeted approach reduced the average uncertainty scores for both sources of uncertainty as opposed to uniform approaches where reductions in average uncertainty from one source came with an increase for the other.

CCS Concepts: • **Human-centered computing** → **Collaborative interaction**.

Additional Key Words and Phrases: crowdsourcing; annotation; ambiguity; calibration

ACM Reference Format:
Quan Ze Chen and Amy X. Zhang. 2023. Judgment Sieve: Reducing Uncertainty in Group Judgments through Interventions Targeting Ambiguity versus Disagreement. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 283 (October 2023), 26 pages. https://doi.org/10.1145/3610074

1 INTRODUCTION

Uncertainty is an unavoidable challenge in many tasks that involve making judgments on items. In particular, judgments that involve groups of people must grapple with uncertainty often, as uncertainty in the group setting can arise from both uncertainty experienced by individuals in the group as well as uncertainty at the group level. Individuals in the group may each feel some level of uncertainty due to the *ambiguity* of the item they are judging, making it hard to personally decide on a judgment. At the same time, even if individuals are certain in their personal judgments, group uncertainty can still arise due to *disagreement* between members of the group, which can come from differences in perspectives of the group members that have not been addressed via a specification in the task instructions.

Authors' addresses: Quan Ze Chen, cqz@cs.washington.edu, University of Washington, Seattle, WA, USA; Amy X. Zhang, axz@cs.washington.edu, University of Washington, Seattle, WA, USA.

2573-0142/2023/10-ART283
https://doi.org/10.1145/3610074

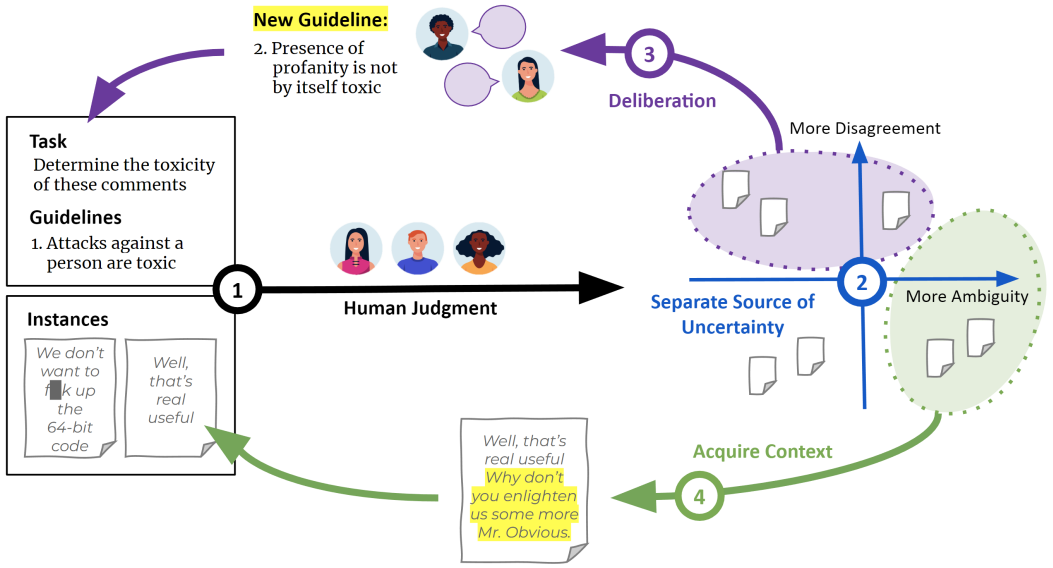


Fig. 1. A high-level overview of the workflow: (1) Human judgments are collected using an annotation tool that quantifies distinct sources of uncertainty; (2) For each instance, scores that correspond to sources of uncertainty (i.e., *ambiguity* and *disagreement*) are computed; (3) Instances with more disagreement are given the DELIBERATION intervention to resolve disagreements, producing new guidelines; (4) Context is collected for instances with more ambiguity and incorporated into the instance.

For example, a group of community moderators determining whether a post should be taken down for being harmful may face uncertainty due to ambiguity in the language used and the poster's intent [66]. At the same time, differences in background and culture [45] mean that members of the community may often disagree on what even is harmful [3] and what actions should be taken as a consequence [2]. Similarly, in the education setting, teaching assistants and instructors are often faced with uncertainty when grading open ended assignments. While the goal of grading is to evaluate a student's level of understanding, a poorly designed assignment question may lead to an answer that does not clearly demonstrate the student's understanding one way or the other. Separately, if a shared grading rubric doesn't specify what to do in a particular case, different graders may end up relying on personal judgment, creating disagreement and inconsistencies [75].

Failure to account for uncertainty during the process of human annotation can lead to unreliable and inconsistent measurements [89] even in domains involving expert judgments [16]. Additionally, biases resulting from different backgrounds and perspectives of individuals in the group can also create biased group judgments if not properly accounted for [71, 72]. Due to these observations, in many areas, approaches and processes have been developed to measure uncertainty in data in order to discard unreliable judgments [6, 34, 67] or to create systems that can make use of disaggregated data [27, 51, 87]. While accounting for existing uncertainty is important for building more robust processes, when it comes to actually making decisions, in many cases intervention to reduce uncertainty becomes necessary.

One intervention to reduce uncertainty tackles the ambiguity in the item being judged by providing additional context to help make a decision. For example, providing information such as the parent post of a comment has been shown to sometimes result in an opposite judgment of the toxicity [66]. However, context is complex, and collecting the full scope of relevant context

for all cases can be difficult [78]. Thus, it is often infeasible to collect context for all instances ahead of time. Another way to reduce uncertainty tackles the disagreement between individual judgments in the group. Several methods have been proposed to address disagreement by adding greater specification to the task, such as through the use of anchoring examples [13] to ground task understanding and using measured uncertainty to find unclear guidelines [56]. Other methods tackle the underlying issue of differing perspectives, where deliberation has been shown to improve consensus [12, 25, 74]. However, these interventions also come at a high cost, often requiring a synchronous collaboration process.

Not only are these interventions costly but applying an intervention meant to address one form of uncertainty when the cause of uncertainty lies elsewhere may lead to wasted effort. For instance, prior work has found that resolution of disagreements via deliberation can fail when the context is ambiguous or missing [74]. On the other hand, more context may not be helpful if group members are certain about their judgment but still need to resolve disagreements. Instead of applying all of these costly interventions in every case that presents uncertainty, if we can measure and distinguish the sources contributing to group uncertainty for each case, then it would be possible to select a more targeted and effective intervention on a per-case level.

In this paper, we present a new workflow, Judgment Sieve, for efficiently reducing uncertainty in group judgments. Judgment Sieveinvolves a decision process that selects a targeted intervention based on the types of uncertainty observed during the initial annotation of each item (Figure 1). When individual uncertainty is detected, we focus on acquiring more context to reduce ambiguity in the item; when disagreement between annotators is detected, we focus on engaging annotators in deliberation to reconcile their diverging perspectives and better specify the task instructions.

We make the following contributions in this paper:

- We present Judgment Sieve, a workflow for reducing uncertainty in group judgment scenarios by utilizing measurements related to ambiguity and disagreement for each instance. We also provide a prototype implementation of this workflow for scalar rating tasks.
- We conduct annotations on two scalar rating task domains: word pair similarity rating (*wordsim*) and toxicity rating (*toxicity*), and verify that adding context and deliberation are **effective interventions for reducing ambiguity and disagreement respectively**.
  - In the top 10% most ambiguous cases, we observed a 21.4% (*wordsim*) and 25.7% (*toxicity*) reduction in ambiguity by introducing context.
  - Similarly for the top 10% highest disagreement cases, we saw a 22.2% (*wordsim*) and 11.2% (*toxicity*) reduction in disagreement by introducing guidelines created from deliberation.
- However, we also observed that a broad application of interventions over all items can increase uncertainty in some circumstance, where adding context increased disagreement by 2.06% (*wordsim*) and 3.54% (*toxicity*).
- We conduct a simulation experiment to evaluate the targeted intervention aspect of Judgment Sieve which selects an intervention based on the type of uncertainty measured in the initial annotation. We find that targeted selection of interventions applied to the most uncertain examples resulted in reductions in the overall means of both uncertainty sources as compared to a uniform approach where reductions in one source of uncertainty came with an increase in the other. Though, we do note that when including instances where our dynamic approach did not assign any intervention, this reduction was not statistically significant.

2 RELATED WORK

There is an increasing recognition in the spaces of machine learning and social computing that accounting for and addressing uncertainty in crowd judgments is an important problem to tackle.

In this section we will review this body of prior work, focusing on: (1) establishing the distinction between error and uncertainty; (2) understanding how (aggregate) measures of uncertainty have been utilized in existing systems and workflows; (3) exploring some theoretical frameworks around distinguishing sources of uncertainty; and (4) discussing prior work around context and deliberation and how they informed the design of the interventions we will be using to address the sources of uncertainty.

2.1 Error v.s. Uncertainty

Uncertainty observed in group judgments in crowdsourced contexts has historically been attributed to *errors*. As a result of the time-constrained nature of crowdsourced tasks and the limited expertise and attention of crowd workers, errors can certainly often occur [44, 76]. Much of the prior work around reducing uncertainty have thus focused on reducing the occurrence and impact of *errors* to the judgment results. One view attributes errors to deficiencies in instructions, leading to confusion and consequent errors on the task [63, 92]. Others have pointed to how better training [29] or selection [54] of workers can reduce error on complex tasks that are difficult to instruct through text and examples alone. Further works have also examined how workers can be kept attentive [4] to maintain quality *throughout* a longer task [37]. From the labor incentive side, some have viewed the problem of *errors* as a result of adversarial intent (i.e., spam) [64, 85], and thus propose incentive-based solutions [22]. Finally, a data (rather than task) focused view poses the idea of utilizing models to correct for errors post-hoc, inferring worker quality from redundant annotations [19] or performance on gold standard questions with known answers [91].

However, this *error*-centric view of uncertainty becomes limited when applied to a more recent demand for human judgment—deciding on cases where there may not be a common understanding of what is *ground truth*. Indeed, applying error reduction mechanisms to complex, socially-situated, and even subjective judgments has resulted in problems with downstream models [32]. Problems like demographic skew [21] or cognitive biases [38] cannot be "patched out" in the same way as errors. In fact, in many such situations, even the introduction of expertise and removal of time constraint does not fully mitigate all uncertainty [1, 24, 65].

The above has led to this work, where we recognize that while *errors* are still a part of observed judgment uncertainty, an increasingly important challenge also falls upon how to account for sources of uncertainty not caused by annotator *error* but rather arise from other factors of the task, like ambiguous cases [5] or legitimate disagreements due to different perspectives and backgrounds of annotators [71]. Consequently, this work doesn't seek to replace traditional error mitigation, but instead serves to improve how we address *non-error* sources of uncertainty.

2.2 Accounting for Uncertainty in Human Judgments

The most straightforward view of uncertainty today, is to treat it as a filter—if the annotators don't agree, then reliable judgment failed. Following this view, solutions focus on measuring and evaluating disagreement, such as through inter-rater reliability metrics [36]. If agreement is too low, data may be discarded, or more annotators brought onboard, until the agreement becomes satisfactory [70, 96]. Achieving a certain threshold of annotator agreement is thus used as a certificate of the quality of a dataset [76]. However, as has been observed, the instances to be judged can themselves lie on a spectrum in terms of how certainly they can be judged [81, 89], so dropping examples or over-sampling annotators could lead to biases that may then be overlooked due to limited transparency in the dataset construction process [31].

Measurement or direct elicitation of uncertainty has also been proposed as a way of building out the task itself. With unclear guidelines or under-specified tasks being collaboratively refined through processes involving the annotators too [9, 11, 46, 55].

One final view on uncertainty focuses on the final application of the human judgment results, proposing that uncertainty be passed along to downstream models as-is. For example, systems have been introduced to learn from uncertain labels [93, 94] and, within the machine learning field, an increasing body of work seeks to harness dis-aggregated labels as the training data for models that achieve higher performance [27, 42]. The application of uncertain labels is not limited to just training, they have also been used in evaluation [34] to produce more realistic assessments of performance.

However, there is likely no one-size-fits-all way to work with uncertainty, and even systems trained with uncertainty data can still fail to be socially cognizant [7]. Not all sources of uncertainty should be treated the same way, even within the same dataset. Our work thus engages with the recent recognition that to optimally address uncertainty, we should categorize and separately quantify it.

2.3 Different Sources of Uncertainty

The quantification of uncertainty has traditionally been done through a statistical lens, by presenting and inferring a probability distribution [17] from the judgments collected. Along this view, recent works have made efforts to quantify uncertainty through capturing *distributions* over the entire set of responses. However, answer distributions are costly to produce [14] and distributions alone don't necessarily offer insight into what may have contributed to the observed result—leading to the need for further analysis and data collection [89].

Thus, more recently, there have been works that aim at *categorizing* different sources of uncertainty separately. Inspired by earlier works in statistics around measurement [39, 48], one such framework proposes the distinction of: *aleatoric* (or aleatory) uncertainty—where uncertainty arises from the natural unpredictable variance in the property/phenomena measured, and *epistemic* uncertainty—where uncertainty arises from a the limitations of our models, tools, and understanding [41, 86]. However, the practical utility of this formulation of uncertainty has sometimes been criticized too, as our evolving understanding of the problem can mean what was once irreducible uncertainty instead was caused by of newly discovered factors [28].

In this work, we don't directly adopt this prior categorization as-is, instead posing a modified categorization by looking at uncertainty through the lens of *ambiguity* and *disagreement*. We elected to use this view as it was more directly aligned to the mechanism of how we collected judgments for our experimental tasks [13] More generally, however, we do also note that our particular formulation of *ambiguity* and *disagreement* isn't meant to be a comprehensive categorization of uncertainty, and depending the judgments and end goals of the task, a different framework for the quantification and classification of uncertainty may be more suitable [77],

2.4 Providing Context to Disambiguate

One source of uncertainty in human judgments is attributed to the *ambiguous* nature of what is being judged. Ambiguity in this sense, could be seen as the result of a lack of sufficient *context* surrounding the case to be judged, and thus an intervention to reduce it lies in adding additional clarity. In toxicity rating tasks, adding context about parent posts has been shown to affect the outcome [66, 88] while context can also be necessary for investigating online abuse cases [59]. Tasks in natural language processing have also seen context added to improve performance [40]. Beyond the limited scope of crowdsourced annotation, context is also commonly used to reduce ambiguity in classical settings like in the legal realm, where judgment processes often involve expert-conducted procedures specifically meant to establish context to remove ambiguity [53]. In education, effective student performance assessment also involves context through free-form responses and intermediate steps [57] in addition to a final answer.

While context is very important, knowing what context to acquire ahead of time, though, can be difficult. For example, in content moderation on Wikipedia, moderators may investigate a variety of factors, such as past behaviors, metadata (like IP addresses), and correlated activity before making a moderation decision [78]. Other communities may choose to investigate different types of context, such as the Civil War portrait identification community drawing upon historic documents and reasoning around timelines as context around the veracity of an identity association [61]. Even classical annotation may demand context, but in this case, it may be additional images to disambiguate occluded areas [52]. As a result, we only want to do extra work to collect uncertainty when we know what is most useful.

2.5 Resolving Disagreement through Rubrics and Deliberation

Another common source of uncertainty in human judgments can be attributed to underlying *disagreements* between individual adjudicators of a case. Many sources can contribute to these disagreements, ranging from inconsistent interpretation of the task criteria to the diverse backgrounds of adjudicators resulting in different perspectives.

For tasks involving crowd annotation, rubrics [30] have been proposed as an effective tool to convey requirements and resolve confusion about aspects of the task. However, rubrics that can cover all the edge cases can be hard to create even by experts, so prior work has utilized crowd participants to help create rubrics [9, 56, 68] by finding areas of high disagreement and asking for suggested guidelines. Rubrics and guidelines can also be implicit, such as in the form of examples [50] or anchors that allow comparison with prior cases [13].

Beyond challenges in creating rubrics, rubrics and guidelines also have limitations when applied. Even when expert-created guidelines are used, adjustments and refinements may be necessary after judgments are made to address issues with the original guidelines [79]. Additionally, beyond layperson crowds, groups of experts judging instances may also just disagree on what the criteria should be [73]. In certain higher-stakes domains like education [75], medical diagnosis [8], or legal judgments [82], it is often the case where the expertise of the humans results in existing guidelines not being applied exactly, instead often conditionally overridden or even contested and overturned. In socially embedded domains, like content moderation, guidelines can also fall out of alignment as distributional properties of the data or adjudicators shifts, such as when social norms shift on online platforms [33, 80] or in broader society. In these situations, past judgments and the criteria that they used may be contested by future adjudicators.

Finally, unclear tasks are not the sole source of disagreement. Even when task goals and guidelines are clear, annotators with different perspective may still disagree about how to judge an item based on different reasoning perspectives [47]. Prior work to automate and scale up deliberation through crowdsourcing has shown that simple reflection-based approaches can be effective at resolving disagreements [23, 49]. More recent work utilizes synchronous deliberation [12, 74] to provide contextual deliberation where those participating in deliberation can quickly form targeted arguments for the particular points of disagreement. Some have also examined the trade-offs between various forms of deliberation design choices (such as the participants, deliberative process, communication medium, etc.) and found that effective deliberation involves building an environment that best matches the task [18]. Of course, more broadly, the successful use of deliberation to resolve disagreement can also depend on other factors. For example, it can be important to make sure deliberation participants are trained to argue effectively [12] and communicate in a way that is collaborative and inclusive [10], especially given the conflicting nature of deliberation. Additionally, the dynamics of deliberation (an intellective task) as groups also means that it can be important to make sure that those matched into deliberation teams are compatible [90].

In our work, we draw from existing literature on disagreement resolution to build out one of our interventions—deliberation. However, designing the right deliberation can be challenging and depends on the participants and tasks, so our application of a simpler form of deliberation may very likely be less effective than customizing deliberation for the tasks involved.

3 DESIGN

In this section we will describe our design of the Judgment Sieve workflow. Our workflow consists of the following procedure (as also illustrated in Fig. 1):

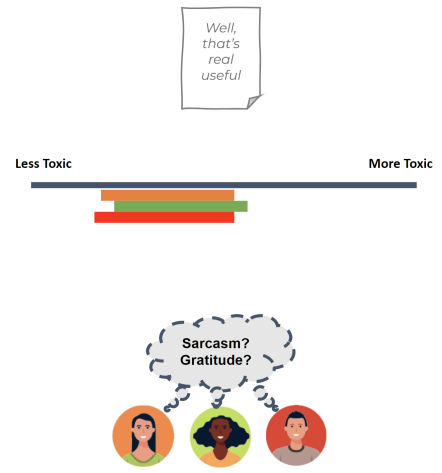
(1) Collect judgments on each instance from individuals in the group using a process that allows measurement of both individual ambiguity and group disagreement for each instance.
(2) Compute two scores for each instance: Ambiguity (M_a) and Disagreement (M_d), based on the measurements in the previous step.
(3) For each instance, based on its ambiguity and disagreement measurements, assign potential interventions:
    - If ambiguity score is above a set threshold, the instance is assigned the CONTEXT intervention. Under this intervention, additional context is gathered for the instance and incorporated into it.
    - If disagreement score is above a set threshold, the instance is assigned the DELIBERATION intervention. Under this intervention, a new group is recruited to re-annotate the instance and then conduct deliberation focusing on their disagreements on the judgment for the instance. At the end of the deliberation for each instance, the group will then collectively produce a suggestion for a new general guideline that they think best resolves the disagreement.
(4) Incorporate the new information produced from the interventions. Additional context acquired is included as part of the corresponding instance. Additional guidelines produced are included into the judgment task definitions.
(5) Repeat the process as necessary until an acceptable level of uncertainty is reached.

In the remaining parts of this section, we will go into more detail about each aspect of the workflow design.

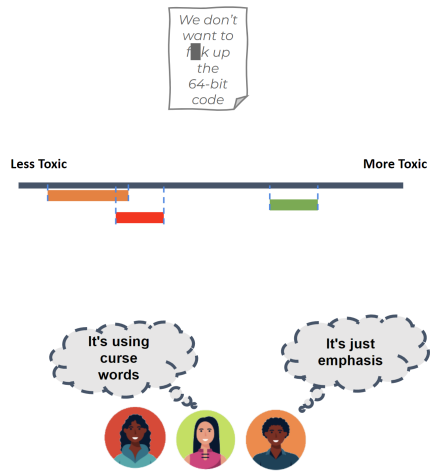
3.1 Measuring Sources of Uncertainty

In order to select the right intervention, we first need an approach to understand what sources may be contributing to the group's current uncertainty on each instance. In our workflow, we focus on two distinct sources of uncertainty: the amount of ambiguity inherent to each judgment (M_a) and the amount of disagreement between judgments (M_d). On a high level, one way to think about the distinction between these two sources of uncertainty is through who contributes to the uncertainty: Ambiguity reflects each individual annotator's certainty about their judgment of the item directly collected through our annotation interface (Fig. 3); Disagreement is an emergent property that results from aggregating judgments across the individual annotators.

In our experiments, we look at a common application of our workflow in the context of making rating judgments on a continuous scale. Before we can apply targeted interventions, we first need an annotation approach that allows us to distinguish different source of uncertainty. Most common annotation tools that focus on rating judgments measure uncertainty through aggregation, utilizing disagreements as a proxy for uncertainty [26, 89]. However, in order to apply effective interventions, we need an approach that is able to distinguish the sources of uncertainty. Some prior work have incorporated means for individual annotators to indicate their confidence through directly providing estimates of their own uncertainty [14], however, humans are generally not good at making these types of assessments [84]. For our specific application of scalar rating, we



(a) When annotators find instances *ambiguous*, wider ranges—reflecting more rating levels they find acceptable—are produced.



(b) When annotators *disagree* about the rating, we will see their ranges be placed in different locations on the scale, resulting in less overlap.

Fig. 2. Illustration showing the how the two sources of uncertainty—ambiguity and disagreement—can manifest in the form of range measurements produced by a range-based rating annotation tool like Goldilocks [13].

make use of an annotation method introduced by prior work, Goldilocks [13], which proposes a way to separately collect measurements on the sources of uncertainty evaluated by each annotator individually during their annotation process. Goldilocks achieves this by adapting rating judgments as a range annotation task where instead of single ratings, raters produce a range ($[l_i^{(x)}, u_i^{(x)}]$) that reflects values that they find acceptable to place the item.

Using the range annotations collected through this approach, we can define two metrics that quantify different sources of uncertainty for each instance (x). We first look at ambiguity—the situation where an individual annotator is unsure about the rating of the instance being judged. With the range-based annotation procedure, we can see that this kind ambiguity would be reflected through the size of the range produced, with "wider" ranges corresponding to more ambiguity around the rating (Figure 2a). Thus we can define an ambiguity score for each instance to be the average size of all ranges collected from the group of annotators participating in the judgment process.

$$\text{Ambiguity}(x, i) = u_i^{(x)} - l_i^{(x)}$$

$$M_a(x) = \frac{1}{|N|} \sum_{i \in N} \text{Ambiguity}(x, i)$$

As for (dis)-agreement between participants, we can see that when range-based annotation is used, the more annotators agree, the more likely it is that the ranges they produce will overlap. So a natural metric can be formed by looking at the amount—in this case the *ratio*—of an annotators range that overlaps with that of another (Figure 2b). However, unlike with range sizes which relate to the fixed scale, simply computing the overlap would result in a metric that is also affected by the size of the ranges (or in our case, the *ambiguity*). We can see that as the absolute size of

any of the ranges increases (reflecting higher *ambiguity*), the likelihood of that range to overlap with another also increases, resulting in a higher overlap ratio. To account for this and derive a metric for disagreement, we don't directly use the overlap ratio, but instead compare the difference between the measured overlap ratio and the *expected* overlap ratio given the size of the ranges being compared. We note that for any range $[l, u]$, the expected overlap ratio of it compared to another uniformly randomly placed range $[l', u']$ is equal to the size of the other range $(u' - l')$. Given the observations above, we define (dis)-agreement as:

$$\text{Overlap}(l, u, l', u') = \max(\min(u, u') - \max(l, l'), 0)/(u - l)$$

$$\text{Agreement}(x, i) = \sum_{j \neq i \in N} \text{Overlap}(l_i, u_i, l_j, u_j) - (u_j - l_j)$$

$$M_{\text{d}}(x) = -\frac{1}{|N|} \sum_{i \in N} \text{Agreement}(x, i)$$

Intuitively, higher agreement scores for an annotator on an instance would indicate more agreement between that annotator and their peers. Positive scores imply that the agreement on this instance was higher than random—that annotators leaned towards *agreement*, while negative scores indicate lower than random agreement—that the annotators leaned towards *disagreement*. We note that such a definition of agreement for each annotator is generally not commutative (i.e., the agreement between a pair of annotators A, B measured from A is not necessarily equivalent to that measured from B). This reflects the natural asymmetry present in agreement as exposed through range overlap—for a hypothetical pair of annotators, the one with a "narrower" (subset) range may agree with their "wider" (superset) partner as both accept the ratings in the "narrow" range, while from the partner's perspective some ratings that they indicated as acceptable were not accepted by their "narrower"-ranged partner. Finally, to make the metric intuitive, we can take the negation of the "agreement" metric to define *dis*agreement. We can arrive at a per-instance disagreement score by taking the average disagreement across all annotators.

3.2 Gathering Additional Context

In cases where ambiguity is high among individual judgments, *context* has been shown to be an effective way to reduce this uncertainty in both traditional human judgment settings [73] as well as for group judgments facilitated in the form of crowdsourced tasks [59]. However, depending on the judgment task involved, *context* itself can encompass a wide variety of types of information, all of which come with varying amounts of cost involved to capture while not necessarily proving effective for reducing ambiguity. Even when capturing context is cheap, presenting too much context can risk exhausting the limited attention capacity of human adjudicators and bog down the judgment process [69] and misleading context could result in bad decisions [83]. As a result, if we want to have human judgments that are scalable, it is likely that attempting to *comprehensively* capture context will be an intractable goal. Thus we need to build a process for gathering additional context that can be informed by measurements of what cases are actually ambiguous and may benefit from context.

In Judgment Sieve, we formulate the collection of context as an open-ended process that is customized depending on the particular domain that our workflow is applied on. While we can't comprehensively provide guidance for all applications, we will give some brief concrete examples of how context may be collected in a couple of setups: one for a more community-involved task of **content moderation**, and one for a more annotation-focused task of **dataset labelling**.

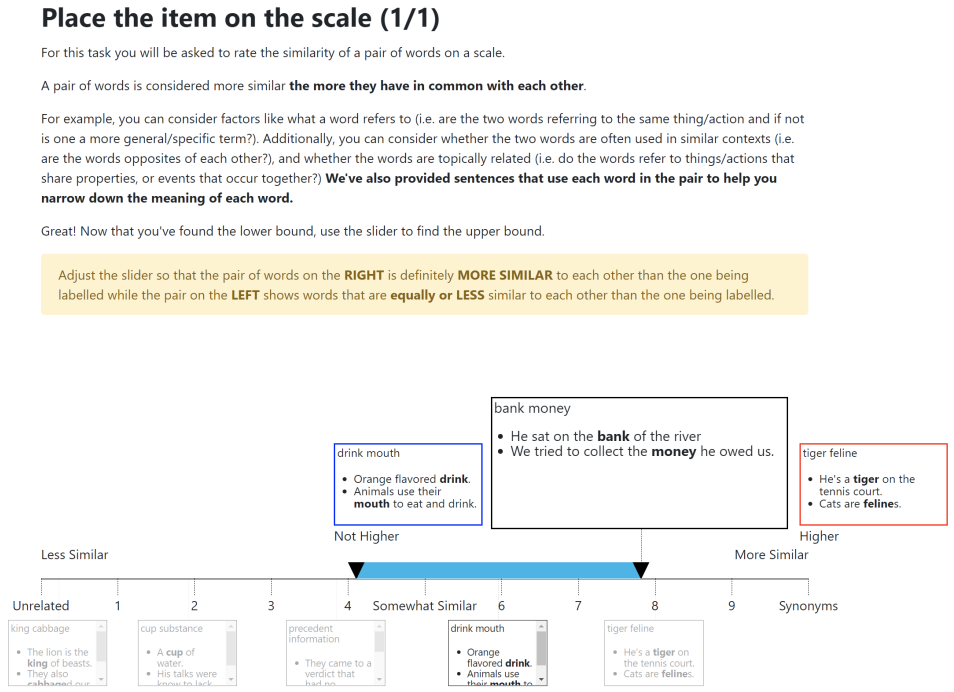


Fig. 3. A screen capture of the interface used in the annotation process. This annotation tool allows us to collect measurements of individual judgments by annotators of their observed ambiguity of each item and allows us to measure disagreement through comparing the ranges across different annotators.

Content moderation in many online communities often takes the form of a committee of *moderators* who collectively decide on a moderation action (such as demoting or removing content, placing a ban on the user, or doing nothing) [25, 58]. While cases may have clear evidence supporting a certain action, historically there have been high-profile cases where limited context contributed to journalistic content being classified as pornographic [33, 43, 62]. For this type of judgment task, Judgment Sieve can identify sets of cases that may require further context due to high ambiguity. The committee can then iterate over such cases to indicate what context might resolve the ambiguity in that case, and deputize a separate group of investigators to collect evidence around the case, who examine logs and metadata and put together a "casebook". This investigation "casebook" would then become the context attached to a case [25]. In the case of **dataset labelling**, Judgment Sieve would directly inform crowd requesters of ambiguous cases through the ambiguity metric. Crowdsourcing requesters might then launch additional tasks, where workers are asked to augment ambiguous instances with hypothetical context that removes the ambiguity, and reincorporating the clarified cases. Alternatively, if data can be re-collected, such if unclear images [20] resulted from their own data collection, requesters might instead trigger the process to collect data again on that instance, and add the new data as context.

In our experiments, we didn't focus on optimizing any specific context gathering strategy. Instead, we simulate a general context acquisition process by taking a dataset that already has contextual cues collected, and withholding the context to mimic a scenario where the efforts to add the context have not yet been made.

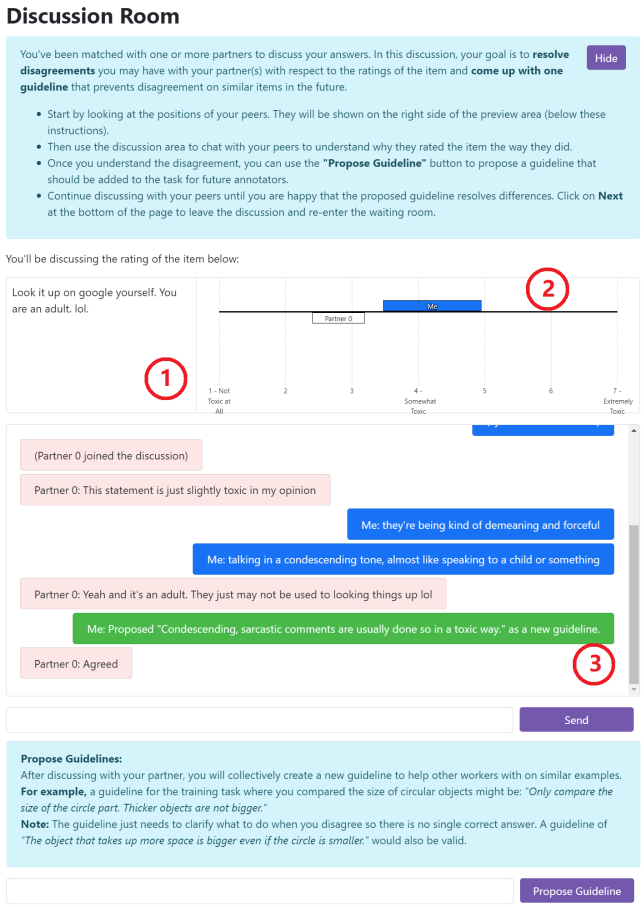


Fig. 4. A screen capture of the deliberation interface used in our experiments. There are 3 main components to the interface: (1) A preview of the instance that was rated, (2) A visualization of the range answers of each participant shown on the same scale, and (3) The synchronous discussion area.

3.3 Using Deliberation to Resolve Disagreement

The design of our deliberation process is inspired by prior work on resolving disagreement using synchronous deliberation [12, 74]. In our workflow, the disagreement metric M_{d} is used to automatically *find* candidate instances that may benefit the most from deliberation—cases where disagreement is the primary source of uncertainty. Then a group deliberates on each example by first independently performing a judgment on the item, and then collectively discussing synchronously. Judgments from each group member is visualized during the deliberation process and the group is collectively prompted to use this to compare their own judgment to those of their peers. Deliberation participants are prompted to consider and elaborate to peers the criteria they used to make their judgment. However, unlike in traditional deliberation systems where the outcome of the deliberation is a judgment on the instance, the goal of our deliberation process is to produce a generalizable guideline for resolving similar disagreements. After engaging in the discussion-based deliberation process, participants are prompted to consider the perspectives they observed during

deliberation as well as the deliberation outcome to collaboratively propose a guideline for future examples that *resolves* the difference in perspective for this instance.

After all the deliberations have concluded, proposed guidelines can be collected and, if needed, de-duplicated. This produces a final set of guidelines that can be incorporated back into the task, so that future disagreements of a similar type are accounted for in the task itself. We note that this overall workflow is also reminiscent of prior methods proposed to utilize worker-provided feedback to improve the quality of instructions in crowdsourcing tasks [56, 68]. However, our workflow makes use of the deliberation process to focus the participants on proposing more effective resolutions that account for the disagreements observed rather than inadequate instructions.

3.4 System Prototype Implementation

In this section we describe some technical details around the prototype[1] that we used to conduct our experiments. To build out the system prototype for our workflow, we created 2 main components: (1) an **annotation** application to collect range-based scalar ratings enabling the measurement of ambiguity and disagreement (Figure 3); and (2) a **deliberation** application that collects range-based ratings and then matches participants into synchronous deliberation sessions (Figure 4).

Our **annotation** application follows the general design of Goldilocks [13], and is implemented as a static web application with the input annotation data and output annotator responses stored directly through Amazon Mechanical Turk (AMT). We use a custom JavaScript toolkit[2] to interface with AMT and coordinate the experiment conditions and data storage. Our **deliberation** application is inspired by the design of prior synchronous deliberation systems [12, 74]. Our front-end interfaces with AMT in a way similar to our annotation application, with the deliberation involving the front-end continuously polling for new messages. We coordinate the matching of participants into synchronous discussion rooms with an additional Python-based back-end that also stores the discussions. The matching of discussion participants is guided by a human operator synchronously monitoring an internal facing dashboard (Section 4.3).

4 EXPERIMENTS

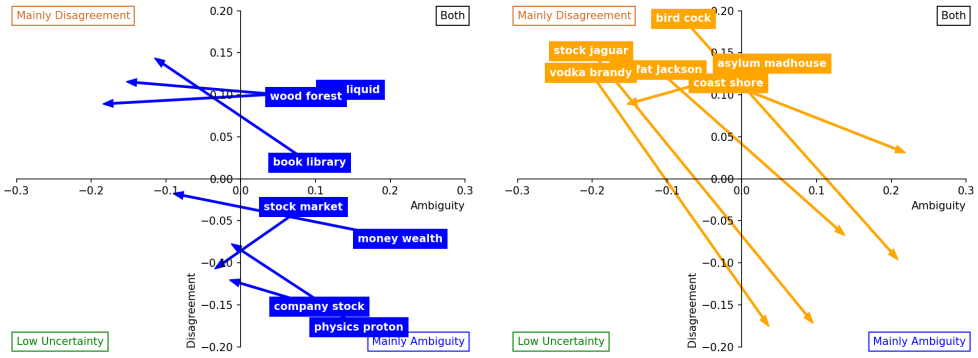
To evaluate the effects of interventions on group judgment uncertainty, we conducted annotation experiments to collect measurements on the uncertainty of group judgments both before any interventions were conducted and after each intervention was applied. For each instance annotated, we collected ambiguity and disagreement measurements using the our range based annotation application under the following conditions: BASELINE—no intervention applied, CONTEXT—context was included as a part of each instance, and DELIBERATION—additional guidelines from the deliberation intervention were provided as part of the task.

4.1 Tasks

For our experiments, we selected two annotation-based tasks that commonly produce uncertainty in group judgments: word similarity (**wordsim**) and toxicity rating (**toxicity**). Both task domains have seen use in prior work and are examples of tasks that contain multiple sources of uncertainty during judgment.

The **wordsim** domain consists of examples based on an the WordSimilarity-535 Test Collection [26] and is structured as a task to judge the relatedness of pairs of words on a 0-10 scale. This domain was selected because it features varied sources that contribute to uncertainty of both the group and individuals. For one, the "relatedness" of words as a concept is only vaguely defined in

[1]Code available: https://github.com/Social-Futures-Lab/targeted-interventions-code
[2]Code available: https://github.com/jmchn1994/amt-shim-template



(a) A sample of items primarily exhibiting ambiguity (blue) and their new uncertainty after applying the CONTEXT intervention.

(b) A sample of items primarily exhibiting disagreement (orange) and their new uncertainty after applying the DELIBERATION intervention.

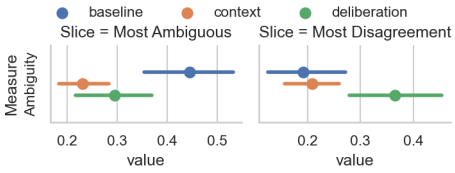
Fig. 5. An illustrated figure showing how the uncertainty of a small sample of items moved within the uncertainty space. Items indicated in orange exhibited primarily disagreement. Items indicated in blue exhibited primarily ambiguity. Arrows point to the new location in the uncertainty space after applying the targeted intervention. Scores are re-scaled such that the origin (0, 0) represents the average ambiguity and average disagreement across all items. Positive values indicate above average uncertainty score measurements.

the **wordsim** task itself, which can lead to different notions of the relatedness between different people reflected as different schools of thought such as comparing the relatedness of words through various facets such as their meaning, usage, generality and occurrence patterns. Additionally, many of the words involved in this task have multiple word senses. Because no context is provided to disambiguate which word sense is implied, individual annotators must also decide how to reconcile the ambiguity resulting from possible word senses. To seed the range-based annotation process, we used the existing similarity annotations to select 5 seed word pair examples that were evenly spaced along the range with the lowest variance. We then assembled our annotation dataset by selecting a random subset of 50 word pairs divided into 5 groups of 10 from the remaining items.

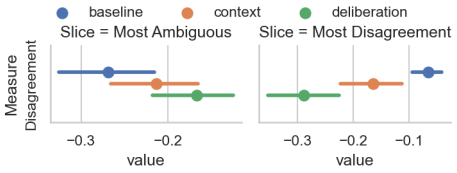
The **toxicity** domain consists of comments collected from a Wikipedia Talk Pages [66] and is structured as a task to judge the toxicity of each individual comment on a continuous rating scale with 7 point semantic differential scale labels. Judging toxicity itself is a task that comes with considerable uncertainty and disagreement. We note that prior work has shown that the background of each annotator and the circumstances in which comments are posted can greatly affect whether the annotator will see the same post as more toxic or not [72]. This gives rise to natural disagreement and ambiguity in annotations. As some comments can be many paragraphs long greatly increasing annotation effort, we first filtered the dataset to select only instances where neither the comment or parent comment exceeded a length of 280 characters. Then, to seed the range-based annotation process, we used the existing toxicity annotations from the dataset source to selected 5 seed comment examples that were evenly spaced along the range with the lowest variance. We then created our annotation dataset by selecting a random subset of 50 comments divided into 5 groups of 10 from the remaining items.

4.2 Acquiring Context

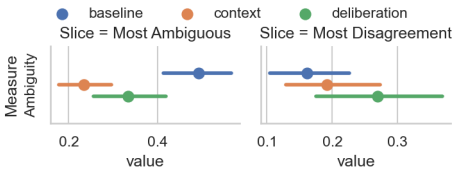
The process of acquiring context in general is usually dependent on the specific goals of the group and the task. As this process is separate from the workflow itself and we did not seek to evaluate



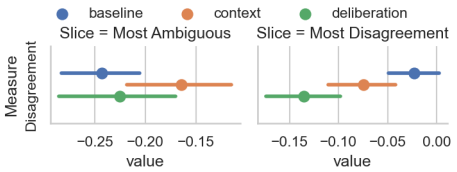
(a) Ambiguity for each slice on the *wordsim* task



(b) Disagreement for each slice on the *wordsim* task



(c) Ambiguity for each slice on the *toxicity* task



(d) Disagreement for each slice on the *toxicity* task

Fig. 6. Point plots for each task domain that shows the ambiguity and disagreement measures under the BASELINE, CONTEXT and DELIBERATION intervention conditions. For each measure, we look at two slices of the dataset: The instances in the top 10% by ambiguity M_a ("Most Ambiguous") and those in the top 10% by disagreement M_d ("Most Disagreement"). Error bars indicate 95% confidence intervals.

the quality of context acquired, we instead simulated the process of acquiring context by using task datasets that were already augmented with context. During annotations in the BASELINE, context of each item was withheld from the annotators, while it was made available during the CONTEXT condition.

For the **wordsim** task, we took inspiration from prior work [40], which used example sentences that contained the word as a way to provide context. For each word in our dataset, we constructed its context by drawing an example sentence that made use of the word in the same form as it appears in the **wordsim** pair. These example sentences were drawn from WordNet [60] when available and when examples were not available, an online dictionary service[3] was used. When multiple word senses existed, a random one was selected to draw the example sentence from. Shorter example sentences were prioritized with long sentences manually simplified. Context for each word pair was then constructed by appending the example sentence for each word involved in the pair.

In the case of the **toxicity** task, our dataset source [66] already contains context information provided in the form of the parent comment of each comment. Context was provided to the annotators by appending the parent comment associated with the item along with a label indicating that it was the parent post.

4.3 Conducting Deliberation

We used a crowd task to conduct deliberation to produce guidelines for the DELIBERATION intervention. At the start of the task, each participant first goes through a training session that teaches them to use the annotation interface. After completing this session, participants are placed in a waiting room where they may be assigned either an assessment session or a deliberation session. In an assessment session, the participant uses the range-based annotation interface to provide their judgment for the instance annotated. In a deliberation session, a participant is matched with 1-2 partners and asked to use a real-time synchronous discussion interface (Figure 4) to discuss the

[3] https://www.merriam-webster.com/

disagreement observed in their range annotations and to collaboratively produce a guideline for future annotators. Guidelines can be proposed or updated by any participant and participants may only leave the discussion after a guideline has been proposed. The allocation of assessment and deliberation sessions was done semi-automatically: While a participant is in the waiting room, the deliberation system makes available a set of sessions available to that participant. A deliberation facilitator can then pick among these options to assign to the participant.

Once the deliberation was complete, the final guideline proposals were collected for each item. We then manually de-duplicated proposals by removing those that were similar. Minor modifications were also made to proposals so that they were phrased in a uniform way for each task domain. The proposals collected were then incorporated into the task instructions for the DELIBERATION condition annotation experiments, with 5 new guidelines added to the **toxicity** task and 6 added to the **wordsim** task.

4.4 Recruitment

We recruited crowd workers from Amazon Mechanical Turk (AMT) to conduct the annotations using an annotation interface based on Goldilocks [13] for each of the conditions: BASELINE, CONTEXT and DELIBERATION. For each condition in each domain, we recruited 25 workers (150 in total). Each participant was given 10 items to annotate for each task deployed. Within each domain, we made sure that a worker could not participate in more than 1 annotation task (displaying a notice and preventing further progression if any tasks beyond the first were attempted), ensuring unique worker pools between conditions in the same task domain. A base payment of $1.0 was given to participants for completing a training task with another $1.0 at the end if they completed all annotations. For each annotation completed, participants were paid $0.3 in the **wordsim** domain ($3.0 total) and $0.5 in the **toxicity** domain ($5.0 total). The median hourly pay was measured to be $13.5 and $15.9 for the two domains respectively.

Additionally, we also recruited separate AMT workers to participate in deliberation sessions on instances in each domain in order to create the guidelines used in the DELIBERATION condition. For each domain and task group, we recruited 4 discussion participants (a total of 40). We used qualifications to ensure that the workers participating in the deliberation sessions did not participate in the annotations. Workers were paid $20 for participating in an hour-long discussion task involving 10 discussion and 10 annotation sessions. A bonus of $4 was given for workers who actively participated in discussions beyond the required 10.

4.5 Simulation Experiment

With the annotation experiment data for each of the 3 conditions collected, we are able to simulate the outcome of selecting a targeted intervention for each instance. For our simulation experiment, we used the ambiguity M_{a} and disagreement M_{d} scores collected during the BASELINE condition to decide the intervention to use for that instance.

For our experiments, we selected a threshold value of 0.1, which targets the instances that ranked in the top 10% in terms of either ambiguity score or disagreement score. To conduct the simulation, instances were sorted by their M_{a} and M_{d} scores collected from the BASELINE condition. We use this to determine a cutoff threshold for the ambiguity and disagreement scores ($\bar{M}_{\mathrm{a}}, \bar{M}_{\mathrm{d}}$). Then, for each instance in the dataset, we first check its ambiguity score. If $M_{\mathrm{a}}(x) \geq \bar{M}_{\mathrm{a}}$, we assign the context intervention by drawing annotation values from the CONTEXT condition for this instance and moving on to the next instance. Otherwise, we check the disagreement score, and if $M_{\mathrm{d}}(x) \geq \bar{M}_{\mathrm{d}}$, we will draw annotation values from the DELIBERATION condition for this instance. If neither uncertainty metric was above the threshold, we leave annotation values from the BASELINE condition unchanged.

4.6 Results

To evaluate our workflow, we focused on 2 main aspects: evaluating the effect of each intervention on the type of uncertainty it targets, and evaluating whether dynamically selecting a targeted intervention based on uncertainty measurements for each example can more efficiently reduce uncertainty compared to a uniform application of intervention.

Specifically, we evaluate the following hypotheses:

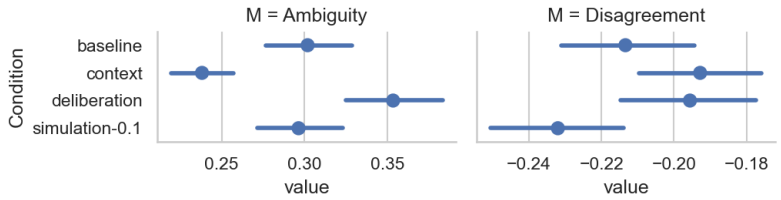
- **H1-a (Interventions are Effective)**: An intervention is effective at reducing the source of uncertainty it targets: CONTEXT will be most effective at reducing ambiguity, while DELIBERATION will be most effective at reducing disagreement.
- **H1-b (Interventions are Targeted)**: An intervention is not effective at reducing the type of uncertainty it does not target.
- **H2 (Efficient Uncertainty Reduction)**: A decision process based only on uncertainty measurements collected without any intervention can select a more optimal intervention for each instance that reduces uncertainty more efficiently than a uniform application of an intervention over all instances.

4.6.1 Effectiveness of Targeted Interventions. In this section, we will examine whether our hypotheses for the effectiveness and targeted nature of interventions is supported in our two task domains. To test our hypotheses, we extract 2 subsets of instances (slices) from each task based on the primary source of uncertainty measured during the BASELINE annotation. For each domain, we selected the top 10% instances that had the highest measured ambiguity as a "Most Ambiguous" slice and the top 10% instances that had the highest measured disagreement as a "Most Disagreement" slice. Then for each set of instances, we tracked their uncertainty after re-annotation following each intervention (CONTEXT and DELIBERATION). We visualize these measurements in Figure 6.

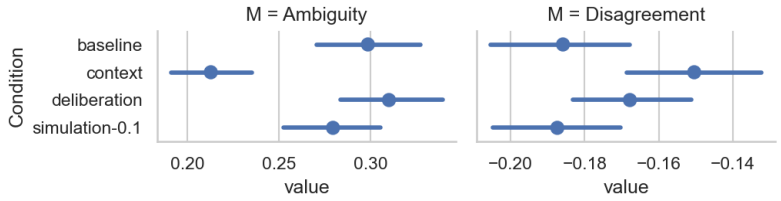
Looking at the slice of "Most Ambiguous" instances in each domain, we found that only the CONTEXT intervention condition was observed to be statistically significant in reducing the ambiguity across both the **wordsim** and **toxicity** task domains ($p < 0.001$, observed only between the BASELINE and CONTEXT conditions using Tukey's HSD). We found similar results for the slice of "Most Disagreement" instances when it came to disagreement, observing only statistically significant reduction in disagreement between DELIBERATION and BASELINE pairings ($p < 0.001$). This supports **H1-a** indicating that interventions are effective in reducing the type of uncertainty it targets.

We also examined how interventions affected the other (non-targeted) source of uncertainty. In both the **wordsim** and **toxicity** domains we did not observe statistically significant interactions of between the non-targeted condition and BASELINE. While lack of observing significance does not indicate that the non-targeted conditions had no effect on the source of uncertainty, it does indicate that they are not as effective as the targeted intervention, thus this provides some partial support for **H1-b**. Curiously, we did find that on the "Most Disagreement" slice in the **wordsim** domain, while DELIBERATION was significant in reducing that disagreement, it also had a significant effect on *increasing* ambiguity. Due to the nature of the task, we hypothesize that the guidelines produced from DELIBERATION resulted in participants considering more factors (word senses, indirect relationships) when determining the relatedness of words and as a consequence of the lack of any other context, they found the instances to be more ambiguous.

4.6.2 Efficiency of Decision Process. Popping up a level and looking at the case of uniformly applying each intervention across the all instances in the entire dataset (Figure 7), we found that CONTEXT was able to reduce ambiguity in both domains ($p = 0.0027 < 0.01$ and $p < 0.001$ for the **wordsim** and **toxicity** domains respectively). However, this seems to also come at a slight cost, also



(a) Comparison for the *wordsim* task domain



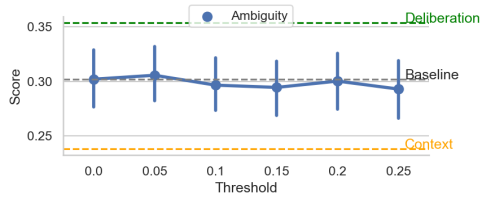
(b) Comparison for the *toxicity* task domain.

Fig. 7. Point plots for each domain that show the ambiguity and disagreement measured after applying a uniform intervention (CONTEXT or DELIBERATION) across all instances and from simulating the selection of different interventions targeted to each instance SIMULATION-0.1. Error bars indicate 95% confidence intervals.

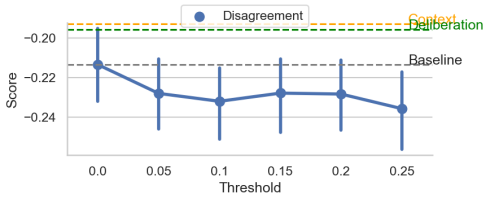
raising the mean disagreement in both cases ($p = 0.026 > 0.01$, not signif.[4], for **toxicity**, $p > 0.01$, not signif., for **wordsim**). This indicates that applying the same intervention across-the-board to all instances can come with trade-offs, potentially causing increases in sources of uncertainty it was not meant to address. When looking at the DELIBERATION condition, we found no statistically significant effects on either uncertainty source when applied across the entire dataset, with slight increases in the mean value on both measurements. This suggests that while deliberation can be useful for instances with the most disagreement, applying it broadly may be harmful. This result is broadly in line with prior work on deliberation that suggests deliberation is likely only effective when items are already low in ambiguity [74] and should be used primarily on the challenging high disagreement cases.

Next, we compare our results from the simulated decision process where instances are assigned different interventions based on whether their uncertainty is primarily caused by ambiguity or disagreement. When comparing against the BASELINE, we found that our simulated process (SIMULATION-0.1) resulted in lower mean values from both ambiguity and disagreement measures in both domains. However, this decrease was not measured to be statistically significant. The lack of significant results is not unexpected, though, as our simulated selection approach only applies an intervention to the top 10% of instances with highest ambiguity and disagreement as measured during the BASELINE annotations (only affecting at most 20% of instances) while all the remaining instances retained their original annotations. We also note that increasing the decision threshold biases results toward the CONTEXT condition—more significant decreases in ambiguity at the cost of higher disagreement. Interestingly, we observed that our two task domains responded differently to our simulated decision process, with **wordsim** achieving the most reduction of uncertainty through reducing disagreement (-8.7%), while **toxicity** achieved more reduction of ambiguity (-6.4%). We

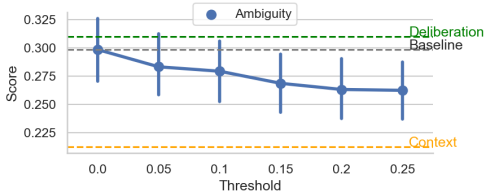
[4]We set an a priori significance level at $p < 0.01$ throughout our statistical tests



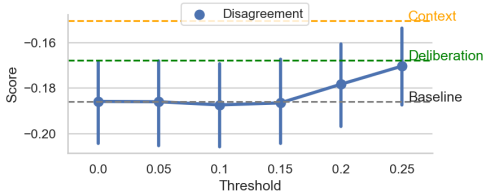
(a) Overall ambiguity on the *wordsim* task



(b) Overall disagreement on the *wordsim* task



(c) Overall ambiguity on the *toxicity* task



(d) Overall disagreement on the *toxicity* task

Fig. 8. Plots showing the simulated interventions applied at different thresholds of 0% (no interventions applied), 5%, 10%, 15%, 20%, and 25%. For all plots, lower values reflect less uncertainty from the corresponding source. Three reference lines are provided on each graph to indicate the average uncertainty measurements of: BASELINE (grey), CONTEXT (orange), and DELIBERATION (green). Error bars indicate 95% confidence intervals around simulations, confidence intervals for the reference lines are not shown (see Figure 7 instead).

hypothesize that this may be due to disagreements being more challenging to resolve in **toxicity** judgments. In the end, while we don't show **H2** to be true in a statistically significant way with one round of targeted intervention, we do see a differences that may allow us to avoid trade-offs of balancing uniformly adding context or deliberating on all instances.

5 DISCUSSION

In this section we will first examine the effect of varying the thresholds for selecting interventions and discuss how thresholds (which affect uncertainty reduction on a per-round basis) work in conjunction with iterative improvement style application of our workflow. Then we will discuss some qualitative observations on the guidelines produced through deliberation and how it may relate to the differences we observe across our two task domains. Following that, we will discuss how our workflow coordinates situations that involve both ambiguity and disagreement and discuss how our workflow can generalize across different tasks and modalities beyond the crowdsourced scalar rating annotation we used in our experiment. Finally, we will discuss some of the limitations of the two interventions we explored—context and deliberation—as well as avenues for future work that may resolve some of these limitations.

5.1 Intervention Selection Thresholds and Iterative Improvement

In section 4.6.2, we found that we are able to observe reductions in both types of uncertainty by simulating a decision process that applied interventions to the top 10% of instances with highest ambiguity and disagreement, respectively, though not at a statistically significant level. As at most 20% of the instances would affected, one question that arises is what happens if we change this threshold to allow interventions to be applied to more (or fewer) instances. To explore this question,

we adjusted the simulation parameters to simulate the decision process under additional thresholds as shown in Figure 8).

Through these simulations, we can observe that the two task domains tested respond differently in terms of their sensitivity to the targeted intervention selected. For the *wordsim* domain, we find that applying targeted interventions reduces *overall* disagreement but achieves relatively little benefit to *overall* ambiguity. From our results in Section 4.6.1, we know that the CONTEXT intervention is effective at reducing ambiguity for those most ambiguous instances, which indicates that the DELIBERATION intervention likely caused increases in ambiguity on the high-disagreement cases that canceled out the reduction of ambiguity provided by CONTEXT. We hypothesize that in this domain, the additional guidelines led to more comprehensive views on "word similarity" with annotators realizing that cases they would have been certain about (and thus disagreed with each other on) were actually ambiguous (and that they wouldn't have considered those alternative interpretations had it not been for the guidelines). On the other hand, for the *toxicity* domain, we find almost the opposite scenario where targeted interventions resulted decreased *overall* ambiguity but had minimal change to (or even increases to) *overall* disagreement. This suggests that for this domain, more context may have reduced the ambiguity around the setting of the online comments, but may have surfaced new disagreements on what toxicity means for the different annotators [72].

While this simulation result itself is interesting, we note that in practice, one would not be able to find an "optimal" threshold using this approach as each intervention would need to be applied to all instances, resulting in an inefficient process. Instead, we posit that optimizing the threshold would likely not be the most effective way to reduce uncertainty in practice; rather, a better approach lies in an **iterative improvement** [35] formulation where our workflow is run in additional iterations that operate on the data and task after application of the interventions from a previous round. Prior work has shown that some uncertainty interventions, like deliberation, used in our workflow may only be effective on instances that have low ambiguity and may be counterproductive otherwise [12, 74]. Indeed, we observe this in Figure 7, where we found that uniformly applying deliberation across all instances can slightly increase overall disagreement in both domains. However, targeted application of deliberation can reduce disagreement even if indiscriminate application does not (Figure 8d). This suggests that a more effective approach lies in iterating on the workflow rather than optimizing thresholds: after each iteration, instances that were ambiguous (and thus not suitable for deliberation) may now be less ambiguous, potentially opening them up to deliberation as an effective intervention in the next round. In an iterative construction of the workflow, selection thresholds can instead be seen as a way to control the rate of uncertainty reduction per-round (almost akin to a "learning rate"). Lower values are more conservative, affecting overall uncertainty less but more likely to avoid interventions cancelling out each others' benefits, whereas higher values reflect a more optimistic view on interventions, increasing the likelihood of failing to reduce uncertainty in a round, but having a larger impact at each step when it works.

5.2 Utility of Guidelines Produced

In our results, we saw that applying deliberation across the entire dataset can result in increases in disagreement even though we also observe that it reduces disagreement for those cases with the highest disagreement. To explore this, we qualitatively examined several of the guidelines produced through the deliberation process to examine how they may not have been effective at scaling to more instances.

For the *wordsim* domain, we found that deliberation resulted in guidelines that outlined additional criteria for what would be considered as "similar", such as: "Antonyms (light/dark, good/evil) are similar.", "Causal [sic] and effect between words make them more similar.", and "Words part of a

natural progression are more similar.". However, while these guidelines would have likely provided more consistent criteria around the word pairs that were deliberated on to produce them, they still leave opportunities for disagreements around applying them—e.g., would a certain word pair be considered a cause-effect pairing or natural progression? For the *toxicity* domain, we found that deliberation resulted in new guidelines such as the following: "Statements about policies not people are not considered toxic.", "Demeaning or condescending statements are likely to be toxic.". Like in *wordsim*, these guidelines are also overall rather narrow ("statements about policies") or could be vague when context was limited ("condescending statements").

While this is not a comprehensive exploration of the effectiveness of producing guidelines, we note that it provides insight into why guidelines produced by our particular deliberation formulation may not have generalized well in some cases. We also note that, even though the setup for Judgment Sieve in our evaluation uses a specific deliberation design, our implementation of it is meant as a proof-of-concept, and doesn't reflect the most effective setup for deliberation optimized for our task domains. In practice, groups utilizing deliberation may elect to use alternative designs that improve matching quality or involve expert-led processes.

5.3 Ambiguity and Disagreement All at Once

As we have observed in our experiments, while ambiguity and disagreement are largely distinct types of uncertainty, it is also not uncommon for an instance to have both high ambiguity and high disagreement. What should one choose to focus on when this occurs? In our simulated version of targeted intervention, we opted to prioritize resolving ambiguity before disagreement. This decision was informed by prior work indicating that the deliberation intervention we used (in the form of self-contained synchronous online discussions) may fail to resolve disagreement in cases with high ambiguity [74]. However, this particular choice may not always be optimal. It may be more productive in some cases to prioritize discussion instead. For example, moderation decisions around developing topics may involve both creating consensus on guidelines (as has been seen in platforms' adaptations to misinformation campaigns related to COVID-19), as well as collecting evidence (current scientific consensus, evaluating whether content is connected to larger misinformation campaigns, etc.) that backs a final decision. In these situations, applying deliberation first or in parallel can shed light on the context that will become necessary, ultimately directing a more effective context collection process. A promising avenue of future work may be to develop approaches to hybridize the collection of context and the deliberation process, allowing groups to switch back-and-forth between the two as needs arise.

5.4 Generalizing our Approach Across Different Tasks and Modalities

In this work, we evaluated Judgment Sieve on the specific judgment modality of continuous scalar ratings for short text-based tasks under a continuous rating scale. However, more broadly speaking, there are many more scenarios (e.g., expert involved group judgments) and judgment modalities (e.g., single or multi-label categorical classification) involving group human judgments where it can be beneficial to reduce uncertainty in a targeted way. In this section, we will discuss two areas where we anticipate opportunities for generalizing our workflow: supporting **complex tasks** through a task specific iterative process, and supporting **modalities beyond scalar rating**.

While Judgment Sieve was built around reducing uncertainty in group judgments in a crowd-sourced setting, the higher level concepts of classifying and quantifying uncertainty and applying targeted interventions could apply to other types of tasks and scenarios. For example, in education settings, a Judgment Sieve-inspired workflow might involve expert-level teaching assistants using a similar tool during grading to measure disagreements around score assignment or ambiguity surrounding some types of answers. These measurements might then feed into group discussions

that result in rubric refinements (to reduce uncertainty) or future updates to assignment questions (to reduce ambiguity). Likewise, online communities may include a Judgment Sieve-inspired process as a part of their moderation process, allowing uncertainty around decisions to audited and addressed as needed. An iterative version of Judgment Sieve for community use might also provide longer term stability as community norms evolve over time [95].

As for other modalities, like categorical or multi-label annotations, Judgment Sieve can be adjusted to use uncertainty categorizations and metrics that are built for those modalities. For example, soft labels [15] could be used in categorical settings, allowing metrics like *dispersion* (e.g., variance) of single-annotator distributions to measure ambiguity and *divergence* (e.g., KL divergence, Wasserstein distance) to measure inter-annotator disagreement. We also do note that, while the potential to generalize is large, Judgment Sieve is still expected to be most effective for reducing uncertainty in more complex or subjective judgment settings, with more perception focused tasks expected to benefit less from a componentized view of uncertainty.

5.5 Caveats of Context

While in general context can reduce ambiguity, in our experiments, we did observe cases where context contributed to an *increase* in uncertainty. These occurred mainly when the context was unexpected. For example, examining the cases where ambiguity increased after context was added in the **wordsim** domain, we saw instances like "bank, money" becoming more ambiguous. However, this did not reflect a failure of the context intervention, rather, we observed that the context introduced (a usage example of "He sat on the **bank** of the river") was likely not a word sense expected by the annotators in the original judgment. Despite increasing uncertainty, this outcome would likely be desirable in practice as it reflected a real increase in uncertainty.

At a higher level, there do exist limits to how far additional context can go to reducing uncertainty in practice and eventually we are likely to run into diminishing returns of gathering more context for minimal benefit to disambiguation. Thus it may be advisable to keep track of the magnitude of uncertainty reduction should further iterations be invoked, and balance the cost accordingly.

5.6 Limits to Scaling Task Specification with Deliberation

Finally, we also note that there are limits to scaling the current design of our deliberation process. In the current process, deliberation produces additional guidelines which are incorporated into the instructions. While processes like de-duplication and reorganization can be done by task requesters, as the task specification becomes increasingly precise, the instructions grounding the task itself can eventually become unwieldy [92]. We see this in current complex tasks like content moderation, where extensive training is given to paid contractors to apply the exact moderation guidelines. If guidelines become too complex, their ability to resolve disagreement would be reduced, as people struggle to understand or find a relevant guideline.

A potential solution to address overly complex task specifications may lie in the realm of legal case building, where decision boundaries can be impossibly complex, and judgments instead often rely on references to past decisions rather than invoking a statutory law or guideline. This opens up possibility of future work to explore alternatives to reducing disagreement beyond deliberation.

6 CONCLUSION

In this paper, we present a new workflow for more efficiently reducing uncertainty in group judgments by applying a targeted intervention on each instance based measurements relating to ambiguity and disagreement. Through our experiments, we find that the interventions of adding context and conducting deliberation do most effectively reduce the type of uncertainty it targets.

We also observe that dynamic selection of interventions on a per-item bases has the potential to avoid the trade-offs in uniformly applying interventions to all items.

REFERENCES

[1] Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM* 2013, 2013 (2013).

[2] Shubham Atreja, Libby Hemphill, and Paul Resnick. 2022. What is the Will of the People? Moderation Preferences for Misinformation. *ArXiv* abs/2202.00799 (2022).

[3] Stephanie Alice Baker, Matthew Wade, and Michael James Walsh. 2020. The challenges of responding to misinformation during a pandemic: content moderation and the limitations of the concept of harm. *Media International Australia* 177 (2020), 103–107.

[4] Michael S. Bernstein, Joel Brandt, Robert C. Miller, and David R. Karger. 2011. Crowds in Two Seconds: Enabling Realtime Crowd-Powered Interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*. Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/2047196.2047201

[5] Lucas Beyer, Olivier J. H'enaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. 2020. Are we done with ImageNet? *ArXiv* abs/2006.07159 (2020).

[6] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. *Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty.* Association for Computing Machinery, New York, NY, USA, 401–413. https://doi.org/10.1145/3461702.3462571

[7] Abeba Birhane. 2021. The Impossibility of Automating Ambiguity. *Artificial Life* 27 (2021), 44–61.

[8] Flora Blangis, Slimane Allali, Jérémie F Cohen, Nathalie Vabres, Catherine Adamsbaum, Caroline Rey-Salmon, Andreas Werner, Yacine Refes, Pauline Adnot, Christèle Gras-Le Guen, et al. 2021. Variations in guidelines for diagnosis of child physical abuse in high-income countries: a systematic review. *JAMA network open* 4, 11 (2021), e2129068–e2129068.

[9] Jonathan Bragg, Mausam, and Daniel S. Weld. 2018. Sprout: Crowd-Powered Task Design for Crowdsourcing. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. Association for Computing Machinery, New York, NY, USA, 165–176. https://doi.org/10.1145/3242587.3242598

[10] Hancheng Cao, Vivian Yang, Victor Chen, Yu Jin Lee, Lydia Stone, N'godjigui Junior Diarrassouba, Mark E. Whiting, and Michael S. Bernstein. 2021. My Team Will Go On: Differentiating High and Low Viability Teams through Team Interaction. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 230 (jan 2021), 27 pages. https://doi.org/10.1145/3432929

[11] 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*. 2334–2346.

[12] Quanze Chen, Jonathan Bragg, Lydia B. Chilton, and Dan S. Weld. 2019. Cicero: Multi-Turn, Contextual Argumentation for Accurate Crowdsourcing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300761

[13] Quan Ze Chen, Daniel S. Weld, and Amy X. Zhang. 2021. Goldilocks: Consistent Crowdsourced Scalar Annotations with Relative Uncertainty. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 335 (oct 2021), 25 pages. https://doi.org/10.1145/3476076

[14] John Joon Young Chung, Jean Y. Song, Sindhu Kutty, Sungsoo (Ray) Hong, Juho Kim, and Walter S. Lasecki. 2019. Efficient Elicitation Approaches to Estimate Collective Crowd Answers. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 62 (nov 2019), 25 pages. https://doi.org/10.1145/3359164

[15] Katherine M. Collins, Umang Bhatt, and Adrian Weller. 2022. Eliciting and Learning with Soft Labels from Every Annotator. In *Proceedings of the Tenth AAAI Conference on Human Computation and Crowdsourcing (HCOMP2022) (HCOMP '22)*. Association for the Advancement of ArtificialIntelligence, Washington, DC, USA.

[16] Corinna Cortes and Neil D. Lawrence. 2021. Inconsistency in Conference Peer Review: Revisiting the 2014 NeurIPS Experiment. *ArXiv* abs/2109.09774 (2021).

[17] Stephen Crowder, Collin Delker, Eric Forrest, and Nevin Martin. 2020. *Introduction to Statistics in Metrology.* Springer.

[18] Todd Davies and Reid Chandler. 2013. Online deliberation design: Choices, criteria, and evidence. *arXiv preprint arXiv:1302.5177* (2013).

[19] A. Philip Dawid and Allan Skene. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of The Royal Statistical Society Series C-applied Statistics* 28 (1979), 20–28.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), 248–255.

[21] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 135–143.

[22] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2012. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *CrowdSearch*.

[23] Ryan Drapeau, Lydia Chilton, Jonathan Bragg, and Daniel Weld. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 4. 32–41.

[24] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Crowdsourcing Ground Truth for Medical Relation Extraction. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 11 (jul 2018), 20 pages. https://doi.org/10.1145/3152889

[25] Jenny Fan and Amy X. Zhang. 2020. Digital Juries: A Civics-Oriented Approach to Platform Governance. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).

[26] Lev Finkelstein, Evgeniy Gabrilovich, Y. Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing Search in Context: The Concept Revisited. *ACM Trans. Inf. Syst.* 20, 1 (jan 2002), 116–131. https://doi.org/10.1145/503104.503110

[27] Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond Black & White: Leveraging Annotator Disagreement via Soft-Label Multi-Task Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2591–2597. https://doi.org/10.18653/v1/2021.naacl-main.204

[28] Craig R Fox and Gülden Ülkümen. 2011. Distinguishing two dimensions of uncertainty. *Fox, Craig R. and Gülden Ülkümen (2011),"Distinguishing Two Dimensions of Uncertainty," in Essays in Judgment and Decision Making, Brun, W., Kirkebøen, G. and Montgomery, H., eds. Oslo: Universitetsforlaget* (2011).

[29] Ujwal Gadiraju, Besnik Fetahu, and Ricardo Kawase. 2015. Training workers for improving performance in crowdsourcing microtasks. In *European Conference on Technology Enhanced Learning*. Springer, 100–114.

[30] 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 (HT '17)*. Association for Computing Machinery, New York, NY, USA, 5–14. https://doi.org/10.1145/3078714.3078715

[31] Timnit Gebru, Jamie H. Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64 (2021), 86–92.

[32] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out?: do machine learning application papers in social computing report where human-labeled training data comes from? *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020).

[33] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

[34] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 388, 14 pages. https://doi.org/10.1145/3411764.3445423

[35] Shinsuke Goto, Toru Ishida, and Donghui Lin. 2016. Understanding Crowdsourcing Workflow: Modeling and Optimizing Iterative and Parallel Processes. In *AAAI Conference on Human Computation & Crowdsourcing*.

[36] Kevin A. Hallgren. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in quantitative methods for psychology* 8 1 (2012), 23–34.

[37] Danula Hettiachchi, Mike Schaekermann, Tristan J. McKinney, and Matthew Lease. 2021. The Challenge of Variable Effort Crowdsourcing and How Visible Gold Can Help. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 332 (oct 2021), 26 pages. https://doi.org/10.1145/3476073

[38] Martin Hilbert. 2012. Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. *Psychological bulletin* 138 2 (2012), 211–37.

[39] Stephen C. Hora. 1996. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety* 54 (1996), 217–223.

[40] Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving Word Representations via Global Context and Multiple Word Prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Jeju Island, Korea, 873–882. https://aclanthology.org/P12-1092

[41] E. Hullermeier and W. Waegeman. 2019. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *arXiv: Learning* (2019).

[42] Oana Inel and Lora Aroyo. 2017. Harnessing Diversity in Crowds and Machines for Better NER Performance. In *ESWC*.

[43] Matthew Ingram. [n. d.]. Here's Why Facebook Removing That Vietnam War Photo Is So Important. *Fortune* ([n. d.]). https://fortune.com/2016/09/09/facebook-napalm-photo-vietnam-war/

[44] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality Management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10)*. Association for Computing Machinery,

New York, NY, USA, 64–67. https://doi.org/10.1145/1837885.1837906
[45] Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R. Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PLoS ONE* 16 (2021).
[46] V K. Chaithanya Manam, Dwarakanath Jampani, Mariam Zaim, Meng-Han Wu, and Alexander J. Quinn. 2019. TaskMate: A Mechanism to Improve the Quality of Instructions in Crowdsourcing. In *Companion Proceedings of The 2019 World Wide Web Conference*. 1121–1130.
[47] Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1637–1648.
[48] Armen Der Kiureghian and O. Ditlevsen. 2009. Aleatory or epistemic? Does it matter? *Structural Safety* 31 (2009), 105–112.
[49] Travis Kriplean, Jonathan T. Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2011. ConsiderIt: Improving Structured Public Deliberation. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11)*. ACM, New York, NY, USA, 1831–1836. https://doi.org/10.1145/1979742.1979869
[50] Ji-Ung Lee, Jan-Christoph Klie, and Iryna Gurevych. 2022. Annotation Curricula to Implicitly Train Non-Expert Annotators. *ArXiv* abs/2106.02382 (2022).
[51] Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 10528–10539. https://doi.org/10.18653/v1/2021.emnlp-main.822
[52] Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. 2018. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE transactions on image processing* 28, 4 (2018), 1575–1590.
[53] E Allan Lind, John Thibaut, and Laurens Walker. 1973. Discovery and presentation of evidence in adversary and nonadversary proceedings. *Michigan Law Review* 71, 6 (1973), 1129–1144.
[54] Angli Liu, Stephen Soderland, Jonathan Bragg, Christopher H. Lin, Xiao Ling, and Daniel S. Weld. 2016. Effective Crowd Annotation for Relation Extraction. In *Proceedings of NAACL and HLT 2016*.
[55] VK Chaithanya Manam and Alexander J Quinn. 2018. Wingit: Efficient refinement of unclear task instructions. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.
[56] V. K. Chaithanya Manam, Dwarakanath Jampani, Mariam Zaim, Meng-Han Wu, and Alexander J. Quinn. 2019. TaskMate: A Mechanism to Improve the Quality of Instructions in Crowdsourcing. *Companion Proceedings of The 2019 World Wide Web Conference* (2019).
[57] Emily Megan Marshman, Ryan Thomas Sayer, Charles Henderson, Edit Yerushalmi, and Chandralekha Singh. 2018. The challenges of changing teaching assistants' grading practices: Requiring students to show evidence of understanding. *Canadian Journal of Physics* 96 (2018), 420–437.
[58] Aiden R. McGillicuddy, Jean-Grégoire Bernard, and Jocelyn Cranefield. 2020. Controlling Bad Behavior in Online Communities: An Examination of Moderation Work. In *International Conference on Interaction Sciences*.
[59] Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. 2021. Abuse is Contextual, What about NLP? The Role of Context in Abusive Language Annotation and Detection. *ArXiv* abs/2103.14916 (2021).
[60] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (nov 1995), 39–41. https://doi.org/10.1145/219717.219748
[61] Vikram Mohanty, David Thames, Sneha Mehta, and Kurt Luther. 2019. Photo Sleuth: Combining Human Expertise and Face Recognition to Identify Historical Portraits. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 547–557. https://doi.org/10.1145/3301275.3302301
[62] Jethro Mullen and Charles Riley. [n. d.]. After outcry, Facebook will reinstate iconic Vietnam War photo. *CNN Business* ([n. d.]). https://money.cnn.com/2016/09/09/technology/facebook-censorship-vietnam-war-photo/index.html
[63] Alexandra Papoutsaki, Hua Guo, Danaë Metaxa-Kakavouli, Connor Gramazio, Jeff Rasley, Wenting Xie, Guan Wang, and Jeff Huang. 2015. Crowdsourcing from Scratch: A Pragmatic Experiment in Data Collection by Novice Requesters. In *HCOMP*.
[64] R. Passonneau and Bob Carpenter. 2013. The Benefits of a Model of Annotation. *Transactions of the Association for Computational Linguistics* 2 (2013), 311–326.
[65] Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics* 7 (2019), 677–694. https://doi.org/10.1162/tacl_a_00293
[66] John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity Detection: Does Context Really Matter?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4296–4305. https://doi.org/10.18653/v1/2020.acl-main.396

[67] Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On Releasing Annotator-Level Labels and Information in Datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 133–138. https://doi.org/10.18653/v1/2021.law-1.14

[68] Vivek Pradhan, Mike Schaekermann, and Matthew Lease. 2021. In Search of Ambiguity: A Three-Stage Workflow Design to Clarify Annotation Guidelines for Crowd Workers. *ArXiv* abs/2112.02255 (2021).

[69] David L. Rosenhan, Sara L. Eisner, and Robert J. Robinson. 1994. Notetaking can aid juror recall. *Law and Human Behavior* 18 (1994), 53–61.

[70] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115 (2015), 211–252.

[71] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *ACL*. https://www.aclweb.org/anthology/P19-1163.pdf

[72] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997* (2021).

[73] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019. Understanding Expert Disagreement in Medical Data Analysis through Structured Adjudication. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 76 (nov 2019), 23 pages. https://doi.org/10.1145/3359178

[74] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 154 (nov 2018), 19 pages. https://doi.org/10.1145/3274423

[75] Arjun Singh, Sergey Karayev, Kevin Gutowski, and Pieter Abbeel. 2017. Gradescope: A Fast, Flexible, and Fair System for Scalable Assessment of Handwritten Work. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale (L@S '17)*. Association for Computing Machinery, New York, NY, USA, 81–88. https://doi.org/10.1145/3051457.3051466

[76] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, 254–263. https://aclanthology.org/D08-1027

[77] Robert Soden, Laura Devendorf, Richmond Y. Wong, Yoko Akama, and Ann Light. 2022. Modes of Uncertainty in HCI. *Found. Trends Hum. Comput. Interact.* 15 (2022), 317–426.

[78] Thamar Solorio, Ragib Hasan, and Mainul Mizan. 2014. Sockpuppet Detection in Wikipedia: A Corpus of Real-World Deceptive Writing for Linking Identities. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, 1355–1358. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1007_Paper.pdf

[79] George Stoica, Emmanouil Antonios Platanios, and Barnab'as P'oczos. 2021. Re-TACRED: Addressing Shortcomings of the TACRED Dataset. In *AAAI Conference on Artificial Intelligence*.

[80] Nicolas Suzor and Darryl Woodford. 2013. Evaluating consent and legitimacy amongst shifting community norms: an EVE Online case study. *Suzor, Nicolas P. & Woodford, Darryl (2013) Evaluating consent and legitimacy amongst shifting community norms: an EVE Online case study. Journal of Virtual Worlds Research* 6, 3 (2013), 1–14.

[81] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 9275–9293. https://doi.org/10.18653/v1/2020.emnlp-main.746

[82] R Peter Terrebonne. 1981. A strictly evolutionary model of common law. *The Journal of Legal Studies* 10, 2 (1981), 397–407.

[83] Craig Thorley, Lara Beaton, Phillip Deguara, Brittany Jerome, Dua Khan, and Kaela Schopp. 2020. Misinformation encountered during a simulated jury deliberation can distort jurors' memory of a trial and bias their verdicts. *Legal and Criminological Psychology* 25 (2020), 150–164.

[84] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185 (1974), 1124–1131.

[85] Jeroen Vuurens, Arjen P de Vries, and Carsten Eickhoff. 2011. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)*. 21–26.

[86] Warren E. Walker, Poul Harremoës, Jan Rotmans, Jeroen P. van der Sluijs, M. B. A. Asselt, Paul Janssen, and Martin Krayer von Krauss. 2003. Defining Uncertainty: A Conceptual Basis for Uncertainty Management in Model-Based Decision Support. *Integrated Assessment* 4 (2003), 5–17.

[87] Dongsheng Wang, Prayag Tiwari, Mohammad Shorfuzzaman, and Ingo Schmitt. 2021. Deep neural learning on weighted datasets utilizing label disagreement from crowdsourcing. *Computer Networks* 196 (2021), 108227. https://doi.org/10.1016/j.comnet.2021.108227

[88] Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*. Association for Computational Linguistics, Austin, Texas, 138–142. https://doi.org/10.18653/v1/W16-5618

[89] Chris Welty, Lora Mois Aroyo, and Praveen Kumar Paritosh. 2019. A Metrological Framework for Evaluating Crowd-powered Instruments. In *HCOMP-2019: AAAI Conference on Human Computation*.

[90] Mark E. Whiting, Allie Blaising, Chloe Barreau, Laura Fiuza, Nik Marda, Melissa Valentine, and Michael S. Bernstein. 2019. Did It Have To End This Way? Understanding The Consistency of Team Fracture. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 209 (nov 2019), 23 pages. https://doi.org/10.1145/3359311

[91] Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. 1999. Development and Use of a Gold-Standard Data Set for Subjectivity Classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, College Park, Maryland, USA, 246–253. https://doi.org/10.3115/1034678.1034721

[92] Meng-Han Wu and Alexander J. Quinn. 2017. Confusing the Crowd: Task Instruction Quality on Amazon Mechanical Turk. In *HCOMP*.

[93] Shuicheng Yan, Huan Wang, Thomas S. Huang, Qiong Yang, and Xiaoou Tang. 2007. Ranking with Uncertain Labels. *2007 IEEE International Conference on Multimedia and Expo* (2007), 96–99.

[94] Hao-Yu Yang, Junling Yang, Yue Pan, Kunlin Cao, Qi Song, Feng Gao, and Youbing Yin. 2019. Learn To Be Uncertain: Leveraging Uncertain Labels In Chest X-rays With Bayesian Neural Networks. In *CVPR Workshops*.

[95] Amy X Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. In *Eleventh International AAAI Conference on Web and Social Media*.

[96] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 35–45. https://doi.org/10.18653/v1/D17-1004

Received July 2022; revised January 2023; accepted March 2023