A Checklist to Combat Cognitive Biases in Crowdsourcing

Tim Draws,[1] **Alisa Rieger,**[1] **Oana Inel,**[1] **Ujwal Gadiraju,**[1] **Nava Tintarev**[2]
[1]Delft University of Technology
[2]Maastricht University
{t.a.draws, a.rieger, o.inel, u.k.gadiraju}@tudelft.nl, n.tintarev@maastrichtuniversity.nl

Abstract

Recent research has demonstrated that cognitive biases such as the confirmation bias or the anchoring effect can negatively affect the quality of crowdsourced data. In practice, however, such biases go unnoticed unless specifically assessed or controlled for. Task requesters need to ensure that task workflow and design choices do not trigger workers' cognitive biases. Moreover, to facilitate the reuse of crowdsourced data collections, practitioners can benefit from understanding whether and which cognitive biases may be associated with the data. To this end, we propose a 12-item checklist adapted from business psychology to combat cognitive biases in crowdsourcing. We demonstrate the practical application of this checklist in a case study on viewpoint annotations for search results. Through a retrospective analysis of relevant crowdsourcing research that has been published at HCOMP in 2018, 2019, and 2020, we show that cognitive biases may often affect crowd workers but are typically not considered as potential sources of poor data quality. The checklist we propose is a practical tool that requesters can use to improve their task designs and appropriately describe potential limitations of collected data. It contributes to a body of efforts towards making human-labeled data more reliable and reusable.

Introduction

Researchers, businesses, and governments use (and reuse) human-labeled data in a wide array of applications, but different types of systemic biases can reduce the quality of this data (Geiger et al. 2020; Faltings et al. 2014). For instance, prominent biases when crowdsourcing data labels include unequal representations of demographic attributes among annotators (Barbosa and Chen 2019) or linguistic biases that lead to stereotypical annotations (Otterbacher 2015).

A relatively less-considered source of poor data quality is *cognitive biases* of crowd workers. Cognitive biases are general human tendencies towards irrationality when making decisions under uncertainty (Tversky and Kahneman 1974). Crowdsourcing is unlikely to be an exception from these tendencies, as crowd workers typically deal with at least some degree of uncertainty with regards to the correctness of the labels they assign. Recent research has indeed shown that cognitive biases such as the *confirmation bias* or *anchoring effect* can negatively affect the quality of crowdsourced annotations (Eickhoff 2018; Hube, Fetahu, and Gadiraju 2019).

Despite this empirical knowledge, crowdsourcing tasks are usually conducted without explicitly considering the influence of crowd workers' cognitive biases on the quality of their annotations. Existing data documentation approaches (e.g., Gebru et al. 2018) aim to make (human-labeled) data sets more reliable by clearly describing the process and purpose of data collection but have so far not included cognitive bias assessments. Moreover, although several methods have been proposed to mitigate cognitive biases in crowdsourcing (Eickhoff 2018; Hube, Fetahu, and Gadiraju 2019), it is currently unclear when different mitigation strategies may be applicable; i.e., there is no protocol by which requesters can identify the specific cognitive biases that may be problematic given a particular task at hand. The large variety and complexity of cognitive biases that have been identified to date (Hilbert 2012) makes this a difficult space to navigate. Requesters need a practical tool that can help them assess which specific cognitive biases may affect crowd workers in a given task at hand so that targeted assessment and mitigation strategies for these biases can be implemented.

In this paper, we propose a 12-item checklist, adapted from business psychology (Kahneman, Lovallo, and Sibony 2011), for combating commonly occurring cognitive biases in crowdsourcing. Each item in this checklist targets a different cognitive bias that may affect crowd workers when labeling data. We explain each bias using a running example of a relevance judgment task and demonstrate the practical application of the checklist through a case study on viewpoint annotations for search results. Finally, by carrying out a large-scale retrospective analysis of relevant studies published at HCOMP over the last three years, we found that cognitive biases apply to a vast majority of crowdsourcing studies but are rarely assessed, accounted for or reported.

All material related to this research (i.e., data sets and analysis code) is publicly available.[1]

Related Work

In this section, we describe previous research on data quality and biases in crowdsourcing. We aim to highlight that, al-

[1]https://osf.io/rbucj

though important advances have been made to make crowdsourced data more accurate and reliable, these approaches usually do not consider the influence of cognitive biases.

Quality in Crowdsourced Annotations

Early research by Snow et al. (2008) showed that crowd workers can perform as well as domain experts in several natural language processing (NLP) tasks such as event temporal ordering, word similarity, and affect recognition. Collecting high-quality annotations from crowd workers, however, is still a challenging task due to concerns posed by identifying, classifying, and counteracting crowd workers' biases and spamming behavior and patterns (Difallah, Demartini, and Cudré-Mauroux 2012; Daniel et al. 2018). Shah, Schwartz, and Hovy (2020) name *label bias* as one of four core sources of bias in NLP models. Several criteria have been shown to influence the quality of crowdsourced annotations, among task and instructions clarity (Kittur et al. 2013; Gadiraju, Yang, and Bozzon 2017; Wu and Quinn 2017), task design (Inel et al. 2018), task difficulty (Mao et al. 2013), incentives (Ho et al. 2015), and quality control mechanisms (Ipeirotis, Provost, and Wang 2010; Dow et al. 2012; McDonnell et al. 2016; Dumitrache et al. 2018).

Following calls for making human-labeled data more reliable (Geiger et al. 2020), several approaches have turned their attention to data documentation; i.e., by tackling issues such as reliability, transparency, and accountability in data collection practices. In the NLP field, Bender and Friedman (2018) proposed *data statements*, a characterization for data sets that provides relevant details regarding the population involved in creating a given data set, how the data set is used in experimental work, and how potential biases in the data set might affect outcomes of the systems that are deployed with it. Gebru et al. (2018) proposed *data sheets* for data sets, a companion document for data sets to exemplify the purpose and composition of the data set, who collected the data and how it was collected, as well as the intended use of the data set. Specifically for crowdsourcing annotations, Ramírez et al. (2020) proposed a set of guidelines for reporting crowdsourcing experiments to better account for reproducibility purposes. Ramírez et al. (2021) then followed up on this work by proposing a checklist that requesters can use to comprehensively report on their crowdsourced data sets. This body of research aligns with and facilitates current efforts towards more trustworthy artificial intelligence through better documentation (Arnold et al. 2019; Stoyanovich and Howe 2019).

Data documentation approaches such as *data statements* (Bender and Friedman 2018) or *data sheets* (Gebru et al. 2018) allow for a thorough assessment for many different types of potential biases (e.g., related to the distribution of crowd workers or the preprocessing of data). However, these methods usually do not consider the influence of *cognitive biases* on data collection.

Cognitive Biases in Crowdsourcing

Cognitive biases are human tendencies towards irrational decision-making or deviation from norms (Tversky and Kahneman 1974). For example, the *confirmation bias* is a tendency to specifically look for information that confirms one's preexisting beliefs (Nickerson 1998). Humans are especially vulnerable to cognitive biases when the cognitive demand of a situation exceeds their currently available cognitive resources (Tversky and Kahneman 1974); e.g. when being confronted with too much or too little information to support a decision or when there is a need to act fast. This can also be the case in crowdsourcing tasks, where objectively "true" answers often do not exist (Aroyo et al. 2019).

Recent research has shown that different types of cognitive biases can negatively impact crowd workers' decision-making and thereby decrease the quality of crowdsourced data labels (Eickhoff 2018; Harris 2019; Hube, Fetahu, and Gadiraju 2019; Saab et al. 2019). For instance, this body of research shows that relevance judgments can be affected by displaying other crowd workers' judgments (i.e., *groupthink* or the *bandwagon effect*) or by revealing information on a single item in subsequent steps (i.e., the *anchoring effect*; Eickhoff 2018). Other work demonstrated that crowd workers may be affected by their personal preexisting attitudes and stereotypes; e.g., when labeling images of faces (Otterbacher et al. 2019) or when judging statements on debated topics (i.e., the *availability bias* and the *confirmation bias*; Hube, Fetahu, and Gadiraju 2019). Furthermore, Gadiraju et al. (2017) found that crowd workers are often unaware of their actual level of competence, which may lead to *overconfidence*. Several strategies have been proposed to assess and mitigate cognitive biases in this context; e.g., by adapting the task design (Barbosa and Chen 2019; Demartini 2019; Eickhoff 2018; Hube, Fetahu, and Gadiraju 2019).

Despite this empirical knowledge of cognitive biases and how to mitigate some of them in the crowdsourcing context, few crowdsourcing studies consider the influence of cognitive biases on data quality — why? Cognitive biases are a vast and complex space that may be hard to navigate for requesters. What is lacking is a practical tool that helps requesters to identify which specific cognitive biases may be problematic in a given task. In practice, such a tool could aid requesters in describing, assessing, and mitigating the influence of the identified potentially problematic cognitive biases. It would contribute to the body of existing efforts towards more reliable and reusable human-labeled data (Gebru et al. 2018; Burmania, Parthasarathy, and Busso 2016).

Introducing a Checklist

Assessing or controlling for cognitive biases in crowdsourcing is currently not straightforward. Identifying which (and how) specific cognitive biases may harm data quality is important but requires a thorough consideration of the task at hand in combination with potentially problematic cognitive biases. However, a plethora of different cognitive biases have been identified to date (Hilbert 2012), and for many cognitive biases, it is still unclear whether or how they may affect crowd workers. This makes navigating the space of cognitive biases extremely complex for requesters. One way to reduce such complexity is to compile a checklist (Gawande 2010).

Kahneman, Lovallo, and Sibony (2011) developed a 12-item checklist for combating cognitive biases in business

decisions. Given a recommended or planned decision, this checklist aims to assist decision-makers in ensuring that their conclusions are as unbiased as possible. Such business decisions may involve, for instance, overhauling a company's pricing structure or acquiring a competitor. Each question in the checklist targets a different cognitive bias (e.g., the *confirmation bias* or *loss aversion*) that may lead to bad decisions in such situations. The 12 items are meant to cover the majority of potential judgment errors that could occur while ensuring that the checklist is concise and easy to use. Although in this case applied to a business context, cognitive biases are general patterns of behavior that humans exhibit when making decisions under uncertainty (Tversky and Kahneman 1974). The basic decision heuristics mentioned in the checklist developed by Kahneman, Lovallo, and Sibony (2011) therefore apply to crowd workers just as well as they do to business decision-makers.

We adapted the checklist developed by Kahneman, Lovallo, and Sibony (2011) to the context of crowdsourcing human-labeled data, by reformulating each of the 12 items to suit the crowdsourcing context. Thus, whereas this adapted checklist practically concerns the same cognitive biases that are mentioned in the original version, it mentions how each of these biases could manifest when conducting a crowdsourcing task. The 12-item checklist we propose is a practical tool that requesters can use to identify potential cognitive biases in the crowdsourcing tasks they design. Each bias in the checklist is accompanied by a guiding question that gives a specific pointer to where it could be applicable. For further illustration, we consider the running example of a simple task in which crowd workers are asked to provide binary relevance judgments on products related to the query "paella pan". We describe the intended use and future development of the checklist in the subsections below.

Cognitive-Biases-in-Crowdsourcing Checklist

1. **Self-interest Bias.** *Does my task offer any room for motivated errors?* That is, could crowd workers have some financial, social, or other self-interest-related incentive to judge particular items differently than others? Crowd workers may (subconsciously) fall prey to self-interest bias due to inadvertent incentives and pricing schemes. For example, if workers receive a financial bonus for each "paella pan"-relevant product they find. Other examples include *social desirability* (i.e., when crowd workers are more likely to make incorrect decisions because other people may examine them; Antin and Shaw 2012) and *satisficing* (i.e., exerting only the minimum required amount of effort into conducting a task to save time or resources; Kapelner and Chandler 2010).
2. **Affect Heuristic.** *Could crowd workers be swayed by the degree to which they 'like' the items they annotate?* For example, crowd workers may be more likely to judge products of a particular brand they like as relevant, independent from the products' true relevance to "paella pan". Phenomena such as *priming effects* (i.e., responding differently depending on a previously presented stimulus) and the *familiarity bias* (i.e., greater favorability towards familiar things or concepts) can play a role here (Morris, Dontcheva, and Gerber 2012).
3. **Groupthink or Bandwagon Effect.** *Does my task design give crowd workers some notion of other people's evaluation of the items they annotate?* For example, crowd workers may judge products as more likely to be relevant to "paella pan" when they see that a majority of other crowd workers have judged this product as being relevant or if it has received high ratings from consumers (Eickhoff 2018).
4. **Salience Bias.** *Could crowd workers' judgments be affected by the salience of particular information?* For example, crowd workers may be more likely to judge products as relevant to "paella pan" if they stand out in an unrelated way (e.g., caps lock titles or high-quality images).
5. **Confirmation Bias.** *Could crowd workers be overly influenced by preconceived notions of the items they annotate?* For example, crowd workers who have a false preexisting idea of what a paella pan is may exhibit confirmation bias if they conduct the task by looking specifically for information that confirms this belief.
6. **Availability Bias.** *Does my task involve judgments related to concepts or people that are likely to elicit stereotypical associations?* For example, crowd workers may be more likely to judge Spanish products as relevant to "paella pan" because they can easily recall numerous examples of the paella dish in Spanish contexts.
7. **Anchoring Effect.** *Is there a possibility that crowd workers overly focus on a specific reference point (i.e., an anchor) when making judgments?* For example, if the first of several products that crowd workers are exposed to are clearly not paella pans (e.g., products unrelated to kitchenware), the first item that somewhat resembles a paella pan (e.g., a regular saucepan) may be more likely to be judged as relevant compared to when the same item was shown in a sequence of actual paella pans. Note that the anchoring effect can also occur within a single human intelligence task (HIT); e.g., when workers are overly influenced by the first information they see (i.e., *primacy effect*), such as the product title, or the last information they see before making their judgment (i.e., *recency effect*).
8. **Halo Effect.** *Does my task involve judgments that could be influenced by irrelevant pieces of information?* For example, crowd workers may be more likely to judge products as relevant to "paella pan" if these products seem suitable for similar dishes (e.g., risotto). This encompasses related biases such as the *decoy effect*, where the choice between two options is affected by the introduction of a (potentially irrelevant) third choice, or the *ambiguity effect*, where (potentially irrelevant) missing information affects crowd workers' decision-making (Eickhoff 2018).
9. **Sunk Cost Fallacy.** *Is the time required to complete my task and what it requires from crowd workers clear at the onset?* The more time and effort crowd workers invest in a task, the more they may want to complete it, despite potentially already having lost interest in the task. This is undesirable as uninterested crowd workers may abandon

a task after investing efforts or complete the task with suboptimal performance (Han et al. 2019b). For example, assuming that crowd workers have to annotate the relevance of 50 different products before completing the task but are not aware of the task length beforehand, their performance may deteriorate in the later stages.

10. **Overconfidence or Optimism Bias.** *Is there a possibility that crowd workers overestimate their ability to perform my task?* For example, it arguably takes a particular level of cooking knowledge to distinguish a paella pan from a regular frying pan or wok. Crowd workers who have never learned about these distinctions may not perceive the task of assigning "paella pan"-relevance judgments to products as hard but may actually not be skilled enough to give high-quality annotations here. This is related to the *Dunning-Kruger effect*, which posits that people with low ability concerning a task tend to be overconfident about their projected performance in it (Kruger and Dunning 1999; Gadiraju et al. 2017).
11. **Disaster Neglect.** *Have crowd workers who commit to my task, been properly informed about the consequences of their participation?* The task selection process is often fairly arbitrary, which means that workers may not realize potential negative effects of committing to a task that they don't have expertise on (Edixhoven et al. 2021). For example, crowd workers may commit to doing "paella pan"-relevance judgments for products on a whim without considering the potential reputation loss and bad annotation quality that could follow if they do not perform well.
12. **Loss Aversion.** *Does my task design give crowd workers a reason to suspect that they may not get paid (fairly) after executing my task?* Due to loss aversion, crowd workers may not select such tasks or abandon them early, leading to a skewed distribution of participants or task starvation (Faradani, Hartmann, and Ipeirotis 2011). For example, if a crowd worker suspects that annotating products in a task will only earn them money if they perform at a particular level, they may abandon the task early to avoid wasting their time and effort (Han et al. 2019a).

How to Use the Proposed Checklist

Here, we give a few pointers regarding the checklist's usage.

When should I apply this checklist? The optimal point to use the checklist is *before data collection*. This allows requesters to not only alert themselves to potential limitations of the data to be collected but also allows for appropriate changes to the task design. If the data have already been collected, requesters may, however, still use the checklist to determine whether cognitive biases may have affected the data in some way (i.e., led to poor data quality or whether the data potentially encodes said biases). The checklist we propose can thus also augment data documentation approaches such as *data sheets* (Gebru et al. 2018).

I applied the checklist to my task design and found at least one potential cognitive bias — now what? The identification of at least one potential cognitive bias in a task design at hand may call for three different actions. First, requesters may want to use this information to *assess* the influence of the identified cognitive biases. The aim behind this would be to check whether these biases truly affect crowd workers during the task. Second, requesters may adapt their task design to *mitigate* the identified cognitive biases. Such adaptations could –at least in some cases– be an easy way to increase data quality without compromising the task design in meaningful ways or vastly elongating the task. Third, especially if data have already been collected from the task at hand, requesters may use the checklist to better *document* their data sets by providing detailed limitations. Pointing out specific cognitive biases that may have affected crowd workers can contribute towards a more accurate data description and thereby make data more reusable. We discuss each of these three actions in more detail below.

How can I *assess* the influence of cognitive biases in my task? Suppose we conclude that our task on relevance judgments for products with respect to the term "paella pan" potentially elicits the *affect heuristic*: we suspect that crowd workers may be more likely to judge products as relevant if they like those products. Previous research suggests that monitoring crowd workers' biases is best done during data collection (Geva, Goldberg, and Berant 2019). We may thus enhance the task design by collecting additional metadata to assess whether crowd workers make erroneous judgments due to the *affect heuristic*. Specifically, we could add an item that measures the degree of crowd workers' personal favorability towards each product they annotate. This would then allow us to approximate the influence of the *affect heuristic* in multiple different ways. For example, we may use a quantitative measure that compares how crowd workers rate items of high and low favorability or conduct a statistical hypothesis test that assesses whether there is a relationship between product favorability and relevance judgments.

How to exactly measure or test for cognitive biases in this context has to be decided individually per suspected bias and the particular crowdsourcing task at hand. To the best of our knowledge, no standard assessments for particular cognitive biases exist in this space. It is nevertheless important to decide on a specific criterion that establishes whether (and perhaps to what degree) bias is present, so that appropriate action can be taken. Below are a few pointers to potential ways of developing such a criterion:

- *Statistical hypothesis tests* are a straightforward way to analyze the presence of systematic patterns in data (e.g., differences between groups or correlations). A caveat of this approach is that failing to reject a null hypothesis may not necessarily mean that no bias is present. That is why we recommend considering not only classical null hypothesis significance testing (i.e., where null hypotheses may be rejected after examining the p-value) but also Bayesian hypothesis testing, which allows for quantification of evidence in favor of either null or alternative hypotheses (Wagenmakers et al. 2018).
- *Self-created or adapted metrics* can be used to quantita-

tively measure patterns or occurrences in data. Here, it is useful to set one or multiple specific thresholds that reflect bias severity before data collection. This can help to decide when the degree of bias is too extreme.

- Statistical techniques such as *structural equation modeling (SEM)* (Ullman and Bentler 2003) or *network analysis* (Epskamp and Fried 2018) may be used to analyze relationships between several factors simultaneously.

It should be pointed out that any such test or metric will only approximate the true, latent influence of the cognitive bias one may wish to assess for. Therefore, we recommend constructing a procedure that consists of several tests and measurements, which build the criterion together. Another useful approach may be to add sanity checks (e.g., by manually evaluating samples of individual cases that show high and low bias according to the criterion). Note also that many statistical procedures (i.e., especially in hypothesis testing) underlie assumptions (Osborne and Waters 2002). For instance, to satisfy the assumption of independence of observations, data may have to be aggregated per crowd worker before conducting a hypothesis test. Requesters should further be aware of common pitfalls in hypothesis testing such as misinterpretation of the *p*-value or statistical power (Greenland et al. 2016).

How can I *mitigate* the influence of cognitive biases in my task? Earlier work has already explored the mitigation of cognitive biases in crowdsourcing tasks. For instance, Eickhoff (2018) showed –through the lens of a standard relevance judgment task– how requesters may deal with biases related to *groupthink*, *anchoring*, and the *halo effect*. Hube, Fetahu, and Gadiraju (2019) investigated how requesters could preempt *confirmation bias* when crowdsourcing subjective judgments related to opinions on debated topics. Next to adapting the task design, requesters may consider improving the data (i.e., the item selection) or changing the worker requirements. Especially difficult tasks may sometimes require non-ambiguous items or particularly qualified workers. Eventually, however, mitigating cognitive biases in crowdsourcing will often require a unique solution that fits the particular task design and suspected cognitive bias at hand. We recommend combining any mitigation efforts with assessments for the suspected cognitive biases to ensure that they have been mitigated successfully.

How can I *document* the influence of cognitive biases in my task? Especially if data have already been collected when applying the checklist, requesters may wish to at least document the potential influence of cognitive biases to make their data more reusable. We recommend augmenting the checklist with general data documentation approaches such as *data sheets* (Gebru et al. 2018). Requesters can add the checklist we propose under a separate section in the data documentation and discuss each bias's potential influence.

Further development and context of this checklist. A few more things should be pointed out to put the usage of this checklist into perspective. First, the checklist –as we propose it in this paper– is unlikely to be exhaustive. We expect that novel research will demonstrate how cognitive biases that we do not yet mention can affect crowd workers. That is why, in contrast to the original checklist developed by Kahneman, Lovallo, and Sibony (2011), we host the latest version of the checklist we propose on an online repository that is open to anyone's contributions.[2] Second, in contrast to our running example, it is unlikely that all of the mentioned biases occur in every crowdsourcing task. We merely posit that any of the 12 mentioned biases could (but do not necessarily do) take place in crowdsourcing. Third, we recommend using more general data documentation approaches such as *data sheets* (Gebru et al. 2018) in tandem with this checklist. Answering questions about the population of crowd workers or the purpose of the (to-be-)collected data set can help distilling potential issues. If the collected data is part of a larger study, we recommend preregistering the research project (Nosek et al. 2018).

Case Study: Viewpoint Annotations for Search Results on Debated Topics

This section demonstrates the practical application of the checklist we propose at the hand of a case study. Our aim in this case study was to collect viewpoint annotations from crowd workers for search results on debated topics.[3] Such data is useful, for example, when aiming to measure the viewpoint diversity in ranked search result lists (Draws et al. 2020; Kulshrestha et al. 2019) or study the effects of viewpoint-biased search result rankings on user attitudes (Draws et al. 2021; Pogacar et al. 2017). We had retrieved search results on nine different debated topics from *Bing*:[4]

1. *Are social networking sites good for our society?*
2. *Should zoos exist?*
3. *Is cell phone radiation safe?*
4. *Should bottled water be banned?*
5. *Is obesity a disease?*
6. *Is Drinking Milk Healthy for Humans?*
7. *Is Homework Beneficial?*
8. *Should People Become Vegetarian?*
9. *Should Students Have to Wear School Uniforms?*

We designed a task wherein crowd workers would be randomly assigned to one of the nine debated topics and see a set of search results related to it. The search results were presented similarly compared to regular search engines (i.e., with a title, snippet, and clickable URL; see Figure 1). Crowd workers would be asked to label each search result for its viewpoint towards the debated topic. Table 1 shows the viewpoint taxonomy we considered: a one-dimensional

[2] See Footnote 1.

[3] We had first collected these data to study user behavior in web search on debated topics (Draws et al. 2021; Rieger et al. 2021).

[4] https://bing.com

representation of the overall stance that a document expresses, ranging from "strongly opposing" to "strongly supporting".[5] Crowd workers would be tasked to annotate the viewpoint of each search result on seven-point Likert scales (i.e., "What stance does this website take on the debated question [*topic*]?"). We also included attention checks between the search results, in which we specifically instructed participants on what option to select on the Likert scale. Full data sheets for the data we collected from this task are available on our repository.[6]

From walking through the checklist before data collection, we could derive that crowd workers' judgments may be affected by three cognitive biases in our task:[7]

1. **Confirmation bias.** We suspected that crowd workers' preexisting attitudes on their assigned topics may affect their annotations. Specifically, we were concerned that crowd workers might interpret their own attitudes into the content they see (especially for ambiguous search results).
2. **Anchoring bias.** Another concern was that crowd workers' first judgment would act as a reference point for the search results to come and thus affect following annotations. Practically, this would have meant that crowd workers' judgments tend towards whatever annotation they gave to the first item they saw.
3. **Halo effect.** We also suspected that crowd workers' preexisting knowledge of their assigned topics may affect their annotations. A halo effect could have occurred if crowd workers have strong preconceived notions about particular subtopics or search result sources, causing them to prematurely rate search results as more extreme (i.e., placing search results into the "opposing camp" or "supporting camp").

We decided to conduct a pilot study to collect annotations for search results on two of the nine debated topics (i.e., *Should zoos exist?* and *Are social networking sites good for our society?*) while assessing the cognitive biases mentioned above. We enhanced our task design with two additional items that collect the necessary contextual metadata. First, to be able to assess the confirmation bias, we measured crowd workers' *personal stance* (i.e., on a seven-point Likert scale ranging from "strongly opposing" to "strongly supporting") on their assigned topic. Second, to enable an assessment of the halo effect, we measured crowd workers' *perceived knowledge* (i.e., on a seven-point Likert scale ranging from "non-existent" to "excellent") of their assigned topic. Assessing the anchoring effect did not require collecting additional metadata.

We published our task on *Amazon Mechanical Turk*[8] to collect viewpoint annotations for all 643 search results that related to the two topics mentioned above. We recruited workers who were located in the United States and had a task approval rate of at least 95%. Crowd workers were paid \$2 for completing the task and could earn a \$0.50 bonus if they clicked on at least half of the links provided in the search results and if they passed both attention checks. We excluded annotations from crowd workers who did not pass at least one of the attention checks.

The data set collected in this pilot study (D_1) contains 1994 annotations from 109 different crowd workers for 643 different search results. Each search result in D_1 pertains to either of the two debated topics *Should Zoos Exist?* or *Are Social Networking Sites Good for Our Society?* and was annotated by two to eleven different crowd workers. Specifically, whereas 92% of search results received three viewpoint annotations, 2% received only two, and 6% received four or more annotations. The low inter-rater reliability between crowd workers who annotated D_1 (Krippendorff's $\alpha = 0.21$) indicates that D_1 contains considerable amounts of noise.[9] Applying the cognitive-biases-in-crowdsourcing checklist and testing for cognitive biases is one way to investigate possible contributing factors to this low data quality.

We conducted several statistical hypothesis tests on D_1 to analyze whether (1) the *confirmation bias*, (2) the *anchoring effect*, or (3) the *halo effect* might have had an influence on crowd workers' annotations.[10]

Confirmation bias. To check whether there was confirmation bias, we conducted classical and Bayesian correlation analyses between crowd workers' pre-existing stance on their assigned topic and their mean viewpoint annotation. We found a significant Spearman correlation ($\rho = 0.27$, $p = 0.002$) and strong evidence in favor of a correlation as part of a Bayesian correlation analysis ($\text{BF}_{10} = 12.49$).[11]

Anchoring effect. We analyzed the influence of a potential anchoring effect by also conducting classical and Bayesian hypothesis tests, this time between crowd workers' first annotation and the mean of their remaining annotations. Here, we found a significant Spearman correlation ($\rho = 0.31$, $p < 0.001$) and extreme evidence in favor of a correlation as part of the Bayesian correlation analysis ($\text{BF}_{10} = 195.33$).

Halo effect. To analyze whether there might have been a halo effect, we compared the range of annotations between crowd workers with lower versus higher knowledge on the topic. We defined "lower knowledge" as low or medium self-reported knowledge on their assigned topic (i.e., the bottom two and central three options on the Likert scale), and

[5] We included two additional options, *neutral* and *irrelevant* (see Figure 1), for search results that did not express any viewpoint or that were found to be irrelevant to the topic, respectively.

[6] See Footnote 1.

[7] Arguably, other biases that are mentioned in the checklist (e.g., the affect heuristic or the availability bias) could have affected crowd workers' judgments as well. For conciseness, however, we keep it to the three biases we mention here.

[8] https://mturk.com

[9] Krippendorff's alpha accounts for missing annotations, so items can vary in terms of how many people annotated them.

[10] The code to replicate all analyses we report here can be found on our repository (see Footnote 1).

[11] We used the R package *BayesFactor* (Morey et al. 2015) to perform Bayesian analyses. We interpret the strength of evidence from Bayes Factors in line with the guidelines proposed by Lee and Wagenmakers (2014), who adapted them from Jeffreys (1939).

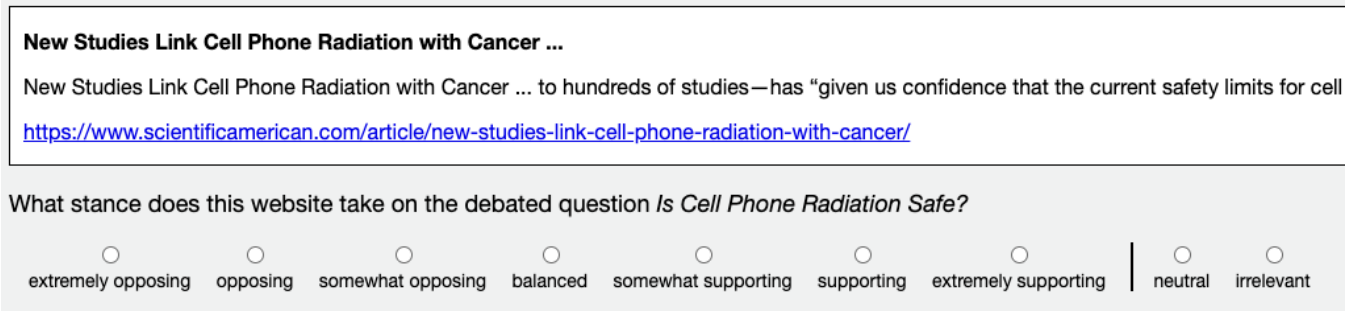

Figure 1: Example item to collect viewpoint annotations for search results in our case study.

l	Label	Example (topic: "Zoos should exist")
-3	strongly opposing	"Horrible places! All zoos should be closed ASAP."
-2	opposing	"We should strive towards closing all zoos."
-1	somewhat opposing	"Despite the benefits of zoos, overall I'm against them."
0	balanced	"These are the main arguments for and against Zoos."
+1	somewhat supporting	"Although zoos are not great, they benefit society."
+2	supporting	"I'm in favor of zoos, let's keep them."
+3	strongly supporting	"There is nothing wrong with zoos – open more!"

Table 1: The viewpoint label taxonomy we considered in our case study. Crowd workers could assign a viewpoint label to each search result by selecting one of the seven options ranging from "strongly opposing" to "strongly supporting". Labels can be represented by integers ranging from -3 to 3 (denoted by l).

"higher knowledge" with those who indicated the top two options on the Likert scale. We did not find any evidence in favor of such a difference as the classical t-test was not significant ($t = 0.42$, $p = 0.68$). A Bayesian t-test revealed moderate evidence in favor of the null hypothesis that there was no difference between these groups ($\text{BF}_{01} = 4.61$).

The analyses we had conducted on D_1 rang the alarm bell for a potential influence of confirmation bias as well as the anchoring effect. We aimed to use this knowledge to inform the collection of labels for search results on all nine debated topics. Further exploratory analyses (e.g., looking at the agreement on different types of search results) led us to suspect that the main source of bias in our crowdsourcing task may have been ambiguous items. Whereas search results that took a clear stance (e.g., "Why Zoos Are An Important Part Of Responsible Wildlife") were often rated quite unanimously, workers diverged when confronted with less strongly opinionated search results (e.g., "Are Zoos Good Or Bad For Animals/Wildlife?"). We thus decided to manually select only those search results that we judged as non-ambiguous or opinionated for our final data collection. Moreover, we suspected that we may have had underestimated the difficulty of the task, so we increased the worker requirements to a HIT approval rate of greater than 98% as well as *Master* status at MTurk.[12] This considerably shrunk our pool of potential crowd workers, so we eased restrictions on workers' location by including other countries where English is spoken as the first or second language by most people (e.g., Australia and Germany). With these changes, we again published our task on Amazon Mechanical Turk.

Selecting only opinionated search results for the nine debated topics resulted in a data set of 480 different search results (approximately balanced over the nine topics). A total of 56 crowd workers provided 1499 viewpoint annotations for this second, final data set of search results (D_2). Each search result pertained to just one of the nine different debated topics and was annotated by three to seven different crowd workers. Inter-rater reliability for this data set is satisfactory (Krippendorff's $\alpha = 0.79$). In contrast to D_1, we did not find evidence for the confirmation bias ($\rho = -0.07$, $p = 0.68$) or the anchoring effect ($\rho = 0.04$, $p = 0.76$) in D_2. Bayesian analyses revealed that the null hypotheses (i.e., that there were no confirmation bias and no anchoring effect) explained the data better than the respective alternative hypotheses ($\text{BF}_{01} = 2.31$ and 3.31, respectively).

Retrospective Analysis of Cognitive Biases in Crowdsourcing

Although the examples and case study we have presented so far relate to specific use cases of crowdsourcing subjective judgments (e.g., relevance judgments), there is reason to expect that cognitive biases occur across different types of crowdsourcing tasks. Cognitive biases are general phenomena that occur when humans make decisions under uncertainty (Tversky and Kahneman 1974) and the checklist we propose covers several different ways in which biases could affect crowd workers (e.g., related to their personal gains, losses, or abilities as well as simple heuristics they may apply while conducting the task). Therefore, in this section, we

[12] Amazon MTurk awards particularly well-performing crowd workers with a *Master* status. This acknowledges the high quality that these workers deliver and allows them to earn higher rewards.

apply our proposed checklist to a set of 27 recent research papers in the crowdsourcing domain. We assess which biases may have been present in the reported studies and whether their (potential) influence was reported upon. By means of this analysis, we aim to show that cognitive biases are often impactful while their influence is not considered in crowdsourcing task designs and publicly available data sets that contain human judgments.

Paper Selection Criteria

We selected research papers for this analysis based on four criteria that all needed to be met for a paper to be included:

1. We selected papers from the 2018, 2019, and 2020 *AAAI Conference on Human Computation and Crowdsourcing* (HCOMP) proceedings, as HCOMP is among the most important venues for research in this area.
2. Papers had to include an online crowdsourcing study in which data was collected (i.e., not using external data).
3. The crowdsourcing task(s) described in the paper had to concern some form of labeling or evaluating data objects in a constrained, closed format (i.e., the crowd workers were given a well-defined answer space). For example, we would include a task that asked crowd workers to judge products as "relevant" or "non-relevant" to the term "paella pan" but we would exclude a task that asked crowd workers to describe products in open text fields.
4. We only included papers in which crowd workers were paid for completing the described task(s).

The selection criteria above were developed among three authors of this paper who acted as independent experts in this study. Using a test sample of 16 papers, the experts ensured that they reached agreement on which papers should be included or excluded based on the four criteria. The final selection procedure resulted in a set of 27 papers (i.e., 4 from 2018, 13 from 2019, and 10 from 2020) that we included in this analysis. We do not report inter-rater reliability, as disagreement between the researchers was resolved through detailed discussions and critical reflection (McDonald, Schoenebeck, and Forte 2019).

Method

Each of the three experts who also co-decided on the inclusion criteria subsequently analyzed each of the 27 selected papers in a two-step process. After reading the paper including the described task design, task instructions, and crowd selection criteria, they first went through each item in the checklist and marked which cognitive bias could have affected the results. Here, the expert would consider the textual description of the task as well as additional available materials (e.g., screenshots). Each bias could be marked with either "yes" (i.e., if there was good reason to assume that the bias may have occurred) or "no" (i.e., if it was impossible or unlikely that the bias had occurred). Therefore, note that a "yes" here did not necessarily mean that crowd workers were indeed affected by the bias, but merely that such an influence could not be ruled out based on the provided task description and additional materials.

As a second step, the expert stated whether the paper at hand discussed the potential influence of cognitive biases on the results. The options here were "yes" (i.e., if the paper identified and at least discussed *all* possible cognitive biases that may have had taken place), "partly" (i.e., if the paper at least discussed a subset of the potential cognitive biases), or "no" (i.e., if the paper did not consider any cognitive biases as a potential influence on the results). Thus, if a paper discussed the potential influence of cognitive biases on the collected data at all, it would receive a "yes" or "partly" label, depending on whether it mentioned all or just a subset of the potential biases identified by the expert. We included this additional label to gauge the degree to which requesters are considerate of such influences on data quality. While it may be difficult to rule out or fully mitigate the influence of cognitive biases on crowdsourcing tasks, discussing potential influences is important information for anyone who may want to use or build on the data set or the published research.

We used majority voting to aggregate the judgments corresponding to the three independent experts. For example, if two of the three experts judged "no" for a particular bias in a particular paper, we would adopt this label for this data point.[13] This resulted in a set of 13 labels per paper (i.e., one for each of the 12 cognitive biases from the checklist as well as the overall judgment on whether the paper considered cognitive biases). Note that this analysis did not concern the methods or evaluations presented in those papers but merely the task design they described.

Results

Table 2 shows the results of our retrospective analysis of crowdsourcing papers at the AAAI HCOMP conference from the last three years. We identified each cognitive bias from the checklist in at least some of the papers we analyzed. Whereas the *salience bias* (93%), *anchoring effect* (81%), and *halo effect* (78%) were marked rather often, biases such as the self-interest bias (30%), loss aversion (22%), or groupthink (15%) were identified comparatively seldom. We also found that some biases often co-occurred in our analysis. Specifically, the *confirmation bias* and *availability bias* as well as *overconfidence* and *disaster neglect* were most often identified for the same papers.

Eight out of the 27 analyzed papers at least partly considered cognitive biases in their task design or discussion. For instance, Otterbacher et al. (2019) show how cognitive biases and stereotypes can affect image labeling and Peng et al. (2019) discuss at length how cognitive biases may affect the hiring process. Mohanty et al. (2019) and Kemmer et al. (2020) acknowledge that a variety of biases such as the confirmation bias can lead to low-quality data labels and propose methods to mitigate these effects.

Note that we intentionally do not disclose which potential cognitive biases had been identified per paper. We wish to point out that this retrospective analysis is not meant to discredit the work of others. Instead, we performed this analysis to show (a) that cognitive biases can occur in a variety of

[13] No conflicts arose for the last (3-option) label as there was a majority judgment for all 27 selected papers.

Bias	2018	2019	2020	Total
Self-interest Bias	0 (0%)	5 (38%)	3 (30%)	8 (30%)
Affect Heuristic	2 (50%)	8 (62%)	5 (50%)	15 (56%)
Groupthink	1 (25%)	2 (15%)	1 (10%)	4 (15%)
Salience Bias	4 (100%)	11 (85%)	10 (100%)	25 (93%)
Confirmation Bias	3 (75%)	8 (62%)	5 (50%)	16 (59%)
Availability Bias	4 (100%)	10 (77%)	5 (50%)	19 (70%)
Anchoring Effect	4 (100%)	9 (69%)	9 (90%)	22 (81%)
Halo Effect	4 (100%)	11 (85%)	6 (60%)	21 (78%)
Sunk Cost Fallacy	3 (75%)	6 (46%)	2 (20%)	11 (41%)
Overconfidence	3 (75%)	9 (69%)	3 (30%)	15 (56%)
Disaster Neglect	1 (25%)	6 (46%)	2 (20%)	9 (33%)
Loss Aversion	0 (0%)	5 (38%)	1 (10%)	6 (22%)
Biases Considered?	1 (25%)	5 (38%)	2 (20%)	8 (30%)

Table 2: Results of the retrospective analysis of cognitive biases in crowdsourcing papers from HCOMP proceedings in 2018, 2019, 2020. Here, *biases considered* refers to papers that discussed the identified cognitive biases at least partly.

crowdsourcing tasks, (b) that the influence of cognitive biases in crowdsourcing is rarely considered, and (c) that the checklist we propose in this paper is widely applicable and could assist researchers in identifying these potential biases.

Discussion

In this paper, we have proposed a 12-item checklist to combat cognitive biases in crowdsourcing. Each item in this checklist refers to a different, commonly occurring cognitive bias that may affect crowd workers' judgments and thereby reduce data quality. Requesters may use our proposed checklist before or after data collection to identify, mitigate, and describe cognitive biases that may influence crowd workers in the tasks they design. To clarify the intended use of the checklist, we have demonstrated its practical application at the hand of a case study on viewpoint annotations for search results. We further showed in a retrospective analysis of recently published crowdsourcing studies that our proposed checklist is widely applicable and that most crowdsourcing studies currently do not consider the influence of cognitive biases on the data labels they obtain.

Limitations

Requesters may apply the checklist we propose to their crowdsourcing tasks but should be aware of at least three important limitations. First, the checklist is unlikely to be exhaustive: several cognitive biases that are relevant to crowdsourcing may still be missing from it. That is why we set up an online repository that will always host the latest version of the checklist and provide an opportunity for contributors to suggest edits. The repository is available at https://osf.io/rbucj. Second, although our proposed checklist can help requesters identify potential cognitive biases that may affect the crowd workers they employ, it does not (yet) give direct recommendations regarding the measurement and mitigation of these biases. We give some pointers in this paper on how the influence of cognitive biases could be assessed or mitigated in certain situations and previous work has already proposed some further mitigation strategies (Eickhoff 2018; Hube, Fetahu, and Gadiraju 2019). However, more research is needed to develop robust procedures that can deal with all the different cognitive biases in our checklist. Third, requesters should be aware that cognitive biases can in some cases be beneficial. *Groupthink*, for instance, is often harnessed to promote collaboration between crowd workers, which can indeed increase data quality (Kobayashi et al. 2018).

Implications

The checklist we propose in this paper has implications for task design as well as data documentation in the crowdsourcing context. As we have discussed, requesters may use this checklist to assess and mitigate cognitive biases. This predominantly concerns adaptations to the task design itself (e.g., adding the collection of contextual metadata) but can also involve item selection or adapting the worker requirements. Furthermore, requesters can use our proposed checklist to document (limitations of) the data they collect. The checklist is applicable to a wide range of crowdsourcing task types, including (but not limited to) validation tasks such as data matching, interpretation and analysis tasks such as relevance judgments, and surveys (e.g., opinion gathering).[14] Although following the procedures we suggest in this paper may increase costs (e.g., due to elongating tasks) and deployment time (e.g., due to prolonged time needed to finetune the tasks), we believe that high data quality and reliability should be any requester's primary aim – especially when the data has a potentially high impact on individuals and society. This is particularly important to facilitate the appropriate reuse of data collections.

Initial steps have been taken towards defining a taxonomy of relevant attributes to report on crowdsourcing studies, such as the employed crowd, the task shown to the workers, the applied quality control mechanisms, and the experimental design (Ramírez et al. 2020; Ramírez et al. 2021). We believe that cognitive biases are an additional factor to consider in reports on crowdsourcing studies. Our retrospective analysis suggests that requesters should also clarify such aspects to the crowd if they aim to mitigate cognitive biases effectively. In particular, the *sunk cost fallacy* could be mitigated by providing the estimated duration of the task in the description of the task. Rejection criteria are also essential for crowd workers in deciding whether they continue to work on a given task (i.e., *loss aversion* and *disaster neglect*). Thus, our retrospective analysis suggests that some aspects that are recommended for reporting on crowdsourcing studies should be included in the actual task design and instructions. This would lead to increased requester-crowd transparency while mitigating several cognitive biases.

Conclusion

Cognitive biases are likely an important limitation of many crowdsourcing studies but are often hard to identify, assess, and mitigate. In this paper, we take one step closer towards tackling cognitive biases in crowdsourcing by proposing a simple, 12-item checklist that requesters can use to decide

[14] We here refer to the taxonomy of microtasks on the web proposed by Gadiraju, Kawase, and Dietze (2014).

whether (and how) some of the most commonly occurring cognitive biases may have an undesired influence in the crowdsourcing tasks they wish to launch or have collected data from in the past. Requesters and researchers can use this tool to improve their task designs and acknowledge cognitive biases as potential sources of sub-optimal data quality where necessary. To this end, we hope that the checklist we propose can contribute towards more reliable, trustworthy, and useful human-labeled data sets.

Acknowledgments

This activity is financed by IBM and the Allowance for Top Consortia for Knowledge and Innovation (TKI's) of the Dutch ministry of economic affairs.

References

Antin, J.; and Shaw, A. 2012. Social desirability bias and self-reports of motivation: a study of Amazon Mechanical Turk in the US and India. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2925–2934.

Arnold, M.; Bellamy, R. K. E.; Hind, M.; Houde, S.; Mehta, S.; Mojsilović, A.; Nair, R.; Ramamurthy, K. N.; Olteanu, A.; Piorkowski, D.; Reimer, D.; Richards, J.; Tsay, J.; and Varshney, K. R. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63(4/5): 6:1–6:13. doi:10.1147/JRD.2019.2942288.

Aroyo, L.; Dixon, L.; Redfield, O.; Rosen, R.; and Thain, N. 2019. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019* 1100–1105. doi:10.1145/3308560.3317083.

Barbosa, N. M.; and Chen, M. 2019. Rehumanized crowdsourcing: a labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Bender, E. M.; and Friedman, B. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6: 587–604.

Burmania, A.; Parthasarathy, S.; and Busso, C. 2016. Increasing the Reliability of Crowdsourcing Evaluations Using Online Quality Assessment. *IEEE Transactions on Affective Computing* 7(4): 374–388. doi:10.1109/TAFFC.2015.2493525.

Daniel, F.; Kucherbaev, P.; Cappiello, C.; Benatallah, B.; and Allahbakhsh, M. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)* 51(1): 1–40.

Demartini, G. 2019. Implicit bias in crowdsourced knowledge graphs. In *Companion Proceedings of The 2019 World Wide Web Conference*, 624–630.

Difallah, D. E.; Demartini, G.; and Cudré-Mauroux, P. 2012. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *CrowdSearch*.

Dow, S.; Kulkarni, A.; Klemmer, S.; and Hartmann, B. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 1013–1022.

Draws, T.; Tintarev, N.; Gadiraju, U.; Bozzon, A.; and Timmermans, B. 2020. Assessing Viewpoint Diversity in Search Results Using Ranking Fairness Metrics. In *Informal Proceedings of the Bias and Fairness in AI Workshop at ECML-PKDD (BIAS 2020)*.

Draws, T.; Tintarev, N.; Gadiraju, U.; Bozzon, A.; and Timmermans, B. 2021. This Is Not What We Ordered : Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380379. doi:10.1145/3404835.3462851.

Dumitrache, A.; Inel, O.; Aroyo, L.; Timmermans, B.; and Welty, C. 2018. CrowdTruth 2.0: Quality metrics for crowdsourcing with disagreement. *arXiv preprint arXiv:1808.06080* .

Edixhoven, T.; Qiu, S.; Kuiper, L.; Dikken, O.; Smitskamp, G.; and Gadiraju, U. 2021. Improving Reactions to Rejection in Crowdsourcing Through Self-Reflection. In *13th ACM Web Science Conference 2021*, 74–83.

Eickhoff, C. 2018. Cognitive biases in crowdsourcing. *WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining* 2018-Febua: 162–170. doi:10.1145/3159652.3159654.

Epskamp, S.; and Fried, E. I. 2018. A tutorial on regularized partial correlation networks. *Psychological methods* 23(4): 617.

Faltings, B.; Jurca, R.; Pu, P.; and Tran, B. D. 2014. Incentives to Counter Bias in Human Computation. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 2(1): 59–66. URL https://ojs.aaai.org/index.php/HCOMP/article/view/13145.

Faradani, S.; Hartmann, B.; and Ipeirotis, P. G. 2011. What's the right price? pricing tasks for finishing on time. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Gadiraju, U.; Fetahu, B.; Kawase, R.; Siehndel, P.; and Dietze, S. 2017. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24(4): 1–26.

Gadiraju, U.; Kawase, R.; and Dietze, S. 2014. A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM conference on Hypertext and social media*, 218–223.

Gadiraju, U.; Yang, J.; and Bozzon, A. 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*, 5–14.

Gawande, A. 2010. *The Checklist Manifesto: How to Get Things Right*. New York: Picador.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Daumé, H.; and Crawford, K. 2018. Datasheets for Datasets URL http://arxiv.org/abs/1803.09010.

Geiger, R. S.; Yu, K.; Yang, Y.; Dai, M.; Qiu, J.; Tang, R.; and Huang, J. 2020. Garbage In , Garbage Out ? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From ? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 325–336. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367. doi:10.1145/3351095.3372862.

Geva, M.; Goldberg, Y.; and Berant, J. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1161–1166. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-1107. URL https://aclanthology.org/D19-1107.

Greenland, S.; Senn, S. J.; Rothman, K. J.; Carlin, J. B.; Poole, C.; Goodman, S. N.; and Altman, D. G. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* 31(4): 337–350. ISSN 15737284. doi:10.1007/s10654-016-0149-3.

Han, L.; Roitero, K.; Gadiraju, U.; Sarasua, C.; Checco, A.; Maddalena, E.; and Demartini, G. 2019a. All those wasted hours: On task abandonment in crowdsourcing. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 321–329.

Han, L.; Roitero, K.; Gadiraju, U.; Sarasua, C.; Checco, A.; Maddalena, E.; and Demartini, G. 2019b. The impact of task abandonment in crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering* .

Harris, C. G. 2019. Detecting cognitive bias in a relevance assessment task using an eye tracker. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 1–5.

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

Ho, C.-J.; Slivkins, A.; Suri, S.; and Vaughan, J. W. 2015. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*, 419–429.

Hube, C.; Fetahu, B.; and Gadiraju, U. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. *Conference on Human Factors in Computing Systems - Proceedings* doi:10.1145/3290605.3300637.

Inel, O.; Haralabopoulos, G.; Li, D.; Van Gysel, C.; Szlávik, Z.; Simperl, E.; Kanoulas, E.; and Aroyo, L. 2018. Studying topical relevance with evidence-based crowdsourcing. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 1253–1262.

Ipeirotis, P. G.; Provost, F.; and Wang, J. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, 64–67.

Jeffreys, H. 1939. *Theory of Probability*. Oxford: Oxford University Press.

Kahneman, D.; Lovallo, D.; and Sibony, O. 2011. Before you make that big decision... *Harvard business review* 89(6). ISSN 00178012.

Kapelner, A.; and Chandler, D. 2010. Preventing satisficing in online surveys. *Proceedings of CrowdConf* .

Kemmer, R.; Yoo, Y.; Escobedo, A.; and Maciejewski, R. 2020. Enhancing Collective Estimates by Aggregating Cardinal and Ordinal Inputs. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8(1): 73–82. URL https://ojs.aaai.org/index.php/HCOMP/article/view/7465.

Kittur, A.; Nickerson, J. V.; Bernstein, M.; Gerber, E.; Shaw, A.; Zimmerman, J.; Lease, M.; and Horton, J. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 1301–1318.

Kobayashi, M.; Morita, H.; Matsubara, M.; Shimizu, N.; and Morishima, A. 2018. An empirical study on short-and long-term effects of self-correction in crowdsourced microtasks. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

Kruger, J.; and Dunning, D. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77(6): 1121.

Kulshrestha, J.; Eslami, M.; Messias, J.; Zafar, M. B.; Ghosh, S.; Gummadi, K. P.; and Karahalios, K. 2019. Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal* 22(1-2): 188–227. ISSN 15737659. doi:10.1007/s10791-018-9341-2.

Lee, M. D.; and Wagenmakers, E. J. 2014. *Bayesian cognitive modeling: A practical course*. Cambridge University Press. ISBN 9781139087759. doi:10.1017/CBO9781139087759.

Mao, A.; Kamar, E.; Chen, Y.; Horvitz, E.; Schwamb, M.; Lintott, C.; and Smith, A. 2013. Volunteering Versus Work for Pay: Incentives and Tradeoffs in Crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 1(1): 94–102. URL https://ojs.aaai.org/index.php/HCOMP/article/view/13075.

McDonald, N.; Schoenebeck, S.; and Forte, A. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW): 1–23.

McDonnell, T.; Lease, M.; Kutlu, M.; and Elsayed, T. 2016. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 4(1): 139–148. URL https://ojs.aaai.org/index.php/HCOMP/article/view/13287.

Mohanty, V.; Abdol-Hamid, K.; Ebersohl, C.; and Luther, K. 2019. Second Opinion: Supporting Last-Mile Person Identification with Crowdsourcing and Face Recognition. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7(1): 86–96. URL https://ojs.aaai.org/index.php/HCOMP/article/view/5272.

Morey, R. D.; Rouder, J. N.; Jamil, T.; and Morey, M. R. D. 2015. Package 'bayesfactor'. *URLh http://cran/r-projectorg/web/packages/BayesFactor/BayesFactor pdf i (accessed 1006 15)* .

Morris, R. R.; Dontcheva, M.; and Gerber, E. M. 2012. Priming for better performance in microtask crowdsourcing environments. *IEEE Internet Computing* 16(5): 13–19.

Nickerson, R. S. 1998. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* 2(2): 175–220.

Nosek, B. A.; Ebersole, C. R.; DeHaven, A. C.; and Mellor, D. T. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences* 115(11): 2600–2606.

Osborne, J. W.; and Waters, E. 2002. Four assumptions of multiple regression that researchers should always test. *Practical assessment, research, and evaluation* 8(1): 2.

Otterbacher, J. 2015. Crowdsourcing stereotypes: Linguistic bias in metadata generated via gwap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1955–1964.

Otterbacher, J.; Barlas, P.; Kleanthous, S.; and Kyriakou, K. 2019. How do we talk about other people? group (un) fairness in natural language image descriptions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7(1): 106–114.

Peng, A.; Nushi, B.; Kıcıman, E.; Inkpen, K.; Suri, S.; and Kamar, E. 2019. What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7(1): 125–134. URL https://ojs.aaai.org/index.php/HCOMP/article/view/5281.

Pogacar, F. A.; Ghenai, A.; Smucker, M. D.; and Clarke, C. L. 2017. The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '17, 209–216. New York, NY, USA: Association for Computing Machinery. ISBN 9781450344906. doi:10.1145/3121050.3121074.

Ramírez, J.; Baez, M.; Casati, F.; Cernuzzi, L.; and Benatallah, B. 2020. DREC: Towards a datasheet for reporting experiments in crowdsourcing. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* 377–382. doi:10.1145/3406865.3418318.

Ramírez, J.; Sayin, B.; Baez, M.; Casati, F.; Cernuzzi, L.; Benatallah, B.; and Demartini, G. 2021. On the state of reporting in crowdsourcing experiments and a checklist to aid current practices. In *Proceedings of the ACM on Human-Computer Interaction (PACM HCI), presented at the 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2021). October 2021*.

Rieger, A.; Draws, T.; Theune, M.; and Tintarev, N. 2021. This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, HT '21, 189–199. New York, NY, USA: Association for Computing Machinery. ISBN 9781450385510. doi:10.1145/3465336.3475101. URL https://doi.org/10.1145/3465336.3475101.

Saab, F.; Elhajj, I. H.; Kayssi, A.; and Chehab, A. 2019. Modelling cognitive bias in crowdsourcing systems. *Cognitive Systems Research* 58: 1–18.

Shah, D. S.; Schwartz, H. A.; and Hovy, D. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5248–5264. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.468. URL https://aclanthology.org/2020.acl-main.468.

Snow, R.; O'connor, B.; Jurafsky, D.; and Ng, A. Y. 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*, 254–263.

Stoyanovich, J.; and Howe, B. 2019. Nutritional labels for data and models. *A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering* 42(3).

Tversky, A.; and Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185: 1124–1131. ISSN 0036-8075, 1095-9203. doi:10.1126/science.185.4157.1124.

Ullman, J. B.; and Bentler, P. M. 2003. Structural equation modeling. *Handbook of psychology* 607–634.

Wagenmakers, E. J.; Marsman, M.; Jamil, T.; Ly, A.; Verhagen, J.; Love, J.; Selker, R.; Gronau, Q. F.; Šmíra, M.; Epskamp, S.; Matzke, D.; Rouder, J. N.; and Morey, R. D. 2018. Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin and Review* 25(1): 35–57. ISSN 15315320. doi:10.3758/s13423-017-1343-3.

Wu, M.-H.; and Quinn, A. 2017. Confusing the Crowd: Task Instruction Quality on Amazon Mechanical Turk. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 5(1): 206–215. URL https://ojs.aaai.org/index.php/HCOMP/article/view/13317.