Position: Insights from Survey Methodology can Improve Training Data

Stephanie Eckman [1] **Barbara Plank** [2 3 4] **Frauke Kreuter** [5 4 1 6]

Abstract

Whether future AI models are fair, trustworthy, and aligned with the public's interests rests in part on our ability to collect accurate data about what we want the models to do. However, collecting high-quality data is difficult, and few AI/ML researchers are trained in data collection methods. Recent research in data-centric AI has show that higher quality training data leads to better performing models, making this the right moment to introduce AI/ML researchers to the field of survey methodology, the science of data collection. We summarize insights from the survey methodology literature and discuss how they can improve the quality of training and feedback data. We also suggest collaborative research ideas into how biases in data collection can be mitigated, making models more accurate and human-centric.

1. Introduction

Social scientists have long relied on survey data collected from human subjects to quantify the population, understand public opinion, and test hypotheses about human behavior. The methods used to collect survey data have been extensively studied and refined by researchers in the field of *survey methodology*, which draws on social and cognitive psychology to develop theories about how humans understand, process, and respond to questions in surveys (Groves et al., 2009).

Data labeled by humans is also central to all stages of the AI pipeline (Plank, 2022; Mazumder et al., 2023), from initial model training, to fine-tuning, reinforcement learning, and model assessment. Insights from social science can contribute to the development of more trustworthy and human-centric models: "if we want to train AI to do what humans want, we need to study humans" (Irving & Askell, 2019). However, collecting high-quality data is difficult, as decades of research in survey methodology and recent high-profile failures in opinion polling (Sturgis et al., 2016; Kennedy et al., 2017; Clinton et al., 2021) demonstrate.

Given the importance of human-labeled data to AI model development, we are surprised that little research in the AI literature has used social science, and survey methodology in particular, to understand the actions and motivations of the humans behind the data generating process. We worry that many researchers collecting data to train, fine-tune, or reinforce AI and ML models are not trained in data collection. A recent paper lamented that, among AI researchers, "everyone wants to do the model work, not the data work" (Sambasivan et al., 2021).

This position paper argues that lessons from survey methodology can improve the quality and efficiency of training data and thus improve models trained on those data. We introduce AI researchers to the community of scientists who want to do the data work and their insights into how to collect high-quality data. We first make the case that label collection is similar to survey data collection (Section 2). Next, we draw on social science theories to develop hypotheses about the facets of the data collection task that may impact the quality of the labels collected (Section 3). Then, we discuss who works as labelers and how the characteristics and uniqueness of the labelers can impact the labels collected and the models trained on those data (Section 4). In Section 5, we join the call for greater transparency in label collection methods, offering lessons from surveys and statistical methods. Throughout the paper, we use the terms labels (and labelers) to refer generally to ML annotations, such as image object labels and bounding boxes, natural language understanding labels, model evaluation data, human feedback for reinforcement learning, and other types of training data (and data generators).

2. How Labeling is Like a Survey

Surveys can be self-administered, like a web or paper-and-pencil survey, or interviewer administered, like a telephone

[1]Social Data Science Center, University of Maryland, College Park, MD, USA [2]Center for Information and Language Processing (CIS), LMU Munich, Germany [3]Computer Science Department, IT University of Copenhagen, Denmark [4]Munich Center for Machine Learning (MCML), LMU Munich, Germany [5]Institute for Statistics, LMU Munich, Germany [6]Joint Program in Survey Methodology, University of Maryland, College Park, MD, USA. Correspondence to: Stephanie Eckman <steph@umd.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

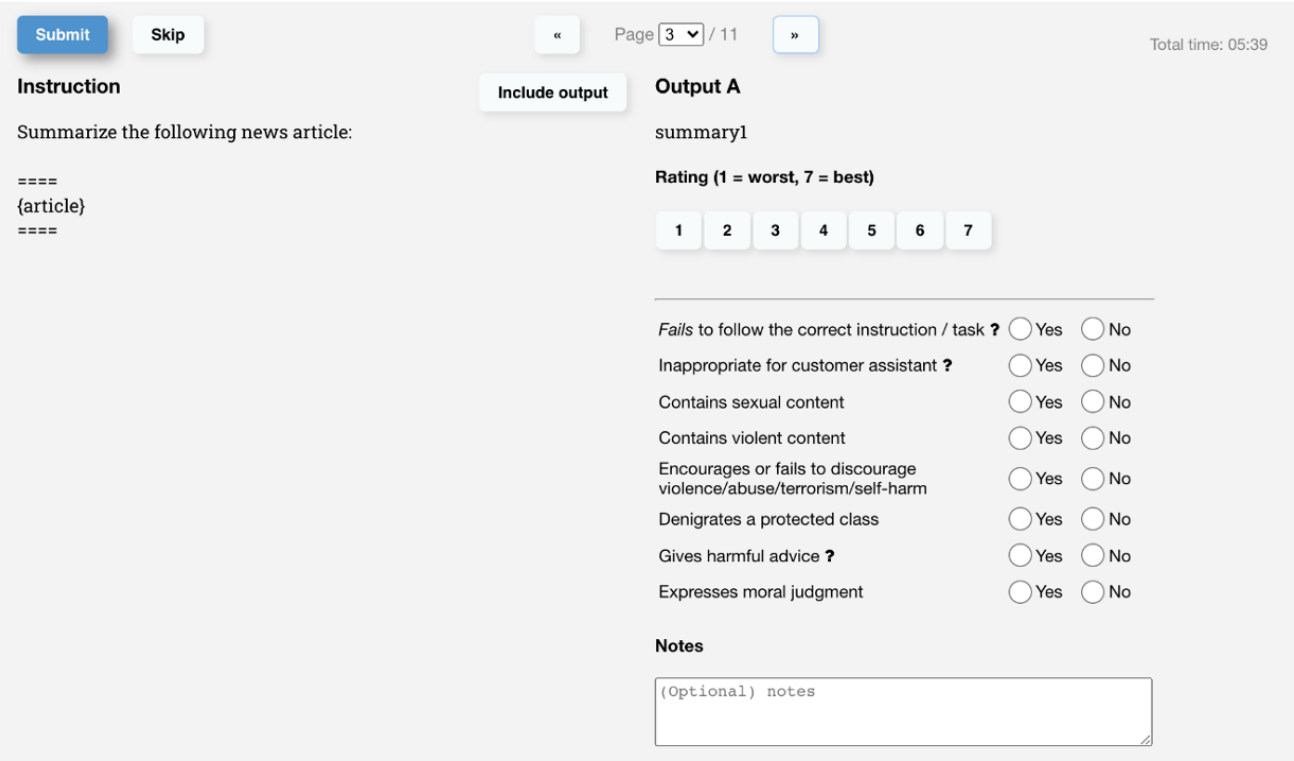


Figure 1. Example of Labeling Interface for InstructGPT (from Ouyang et al., 2022)

or face-to-face survey. A web survey presents a series of questions with, most often, closed answer choices. Superficially, the process of labeling observations for ML models often looks like a web survey: labelers see one or more prompts and associated answer choices (Figure 1). As in surveys, the task can be factual (is there a bicycle in this photo?) or opinion based (which of these responses to the prompt is most helpful?).

Of course, label collection differs from survey data collection in important ways. Surveys ask many different questions and end when the questions run out; labeling tasks usually consist of similar repeated observations, and labelers often continue until they choose to stop. Surveys usually ask people about themselves: their opinions, behaviors, characteristics. Labeling tasks more often involve passing judgment on an object outside of oneself (for example, images, product reviews, or a news article). However, the human feedback data used in reinforcement learning often aims to capture personal opinions: "which of the following responses is most relevant to the prompt?"

The goal of surveys and labeling tasks is also different. Surveys ask questions of a sample of selected persons to make inference about the population. For example, say we wished to know what proportion of the U.S. adult population does not have health insurance. We select a sample of U.S. adults and ask them whether they have health insurance. For each respondent, we have a yes or no response to the question.[1] The proportion of respondents without insurance in the sample is an estimate of the proportion of the population without insurance. A survey often asks many questions on related topics. Our example survey might also ask "when was the last time you visited a health care professional?" and "have you received a flu vaccine in the last 12 months?" Results are often reported at the question level although relationships between questions are also of interest: for example, are people with health insurance more likely to get a flu vaccine? The quality of the analysis rests on collecting *accurate responses* from a *representative sample*.

The goal when collecting labeled data is not to estimate the proportion of the population that finds a given post offensive or detects a vehicle in a given image. Instead, the goal is to learn patterns from the labeled data to predict labels for unseen observations. Thus, while it is not as important that labels reflect the population's views at the observation level, across observations, the views of the population should be represented. Collecting *accurate labels* from a *diverse set of labelers* is important to the performance and generalizability of the final model.

Despite their differences, both surveys and label collection

[1]We are ignoring "I don't know" and other types of responses.

have these two needs in common. The data points provided in response to a prompt should capture the data provider's judgment: accurate responses / accurate labels. Those who provide the data points should represent the judgments of the relevant population: representative sample / diverse set of labelers. Sections 3 and 4 discuss these needs in turn.

3. Need for Accuracy

Survey methodologists have developed theories of how respondents understand, process, and respond to survey questions. We summarize these theories and use them to derive hypotheses about the aspects of the label collection task that may impact label quality. We end the section with thoughts on mitigation measures and future research.

3.1. Response Process

Ideally, survey respondents understand questions thoroughly and respond thoughtfully. However, they make take shortcuts that can threaten data quality.

Optimal Survey Response Process Responding to a survey question can involve several cognitive steps (Tourangeau et al., 2000; Tourangeau, 2018):

1. Comprehension: Understand the question and the response options
2. Retrieval: Search memory for relevant information
3. Integration: Integrate the retrieved information to form an answer to the question
4. Mapping: Map that answer onto the provided answer choices

Ideally, a respondent proceeds through each step in order. However, they can choose to backtrack. For example, considering the response options in the Mapping step may change the interpretation of the question (Comprehension) or trigger additional relevant information to come to mind (Retrieval).

The above model exposes why respondents sometimes give incorrect answers. At the Comprehension step, they may fail to understand the question or some of the words it uses. They may have a different understanding of some of the words than those who wrote the question. At the second step, respondents may fail to retrieve all relevant information. Some information may have been forgotten. At the third step, respondents may fail to put in the mental effort to bring together their understanding of the question with the retrieved information. At the fourth step, respondents may not find an answer choice that reflects their answer or they may edit the true answer to avoid revealing sensitive information.

Deviations from Optimal Response Process The full survey response process outlined above is cognitively demanding. Some respondents resort to taking shortcuts, an approach called satisficing (Krosnick, 1991; Krosnick et al., 1996). For example, they may retrieve only the most recent relevant information from memory (recency bias) or choose the first reasonably correct answer choice. Satisficing relates to the cognitive miser theory in psychology, which holds that people seek to minimize cognitive effort (Fiske & Taylor, 1991; Kahneman, 2011).

As predicted by satisficing theory, eye-tracking studies show that respondents do not read all options in *select-all-that--apply* questions (Galesic et al., 2008), and shortcuts are more common as survey length increases (Galesic & Bosnjak, 2009). Respondents tend not to read provided instructions (Brosnan et al., 2019) or click on provided definitions (Peytchev et al., 2010), especially when they believe they understand the concept that is asked about (Tourangeau et al., 2006).

The survey literature discusses several types of more extreme undesirable response behavior. Acquiescence is the tendency to say "yes" to *yes/no* questions, regardless of content (Knowles & Condon, 1999). Straightlining is the practice of choosing the same response option in the same position (for example, the first response option) to all questions. This behavior is most common in batteries or grids of questions with the same response options (Kim et al., 2019). Some respondents even deliberately give incorrect answers to later questions to reduce the length or burden of a survey: when a "yes" response triggers follow up questions, respondents may learn to report "no." This phenomenon is called motivated misreporting (Kreuter et al., 2011; Tourangeau et al., 2012; Eckman et al., 2014).

Context Effects Perceptions and judgments are shaped by the broader context and preceding experiences, a phenomenon called context effects (Tversky & Kahneman, 1974; Strack, 1992). For example, a very tall person can make others seem shorter: a contrast effect. An unethical politician can make other politicians seem less ethical: an assimilation effect (Bless & Schwarz, 2010).

Opinion questions are especially vulnerable to context effects, because respondents do not always have well-formed, fixed opinions that they retrieve from memory. Instead, they form opinions when asked for them, and this process can be shaped by context clues in the question, the response options, the look and feel of the instrument, or the previous questions (McFarland, 1981; Zaller & Feldman, 1992; Schwarz, 2007).

Order effects are the most common example of context effects in surveys. Questions that come earlier in a survey can change how respondents interpret later questions. Re-

searchers have found order effects in reports of crime and bullying victimizations, of disabilities, and of race (Cowan et al., 1978; Gibson et al., 1978; Bates et al., 1995; Todorov, 2000; Huang & Cornell, 2015).

3.2. Hypotheses about Label Quality

These theories about how respondents answer questions lead us to several hypotheses about the properties of the labeling task that may impact training data quality. The ML literature has investigated some of these hypotheses, but fundamental research gaps exist.

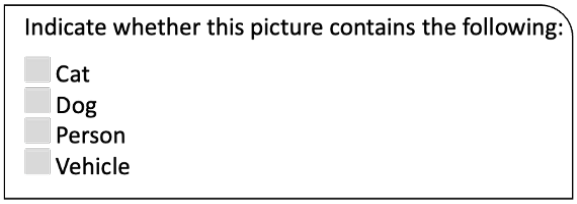
Wording and Reading Level Respondents cannot provide high-quality answers to questions unless they clearly understand what the question asks and what they should include and exclude in their answer. Questions should be at an eighth grade reading level or lower (Dillman et al., 2014) and terms should be as unambiguous as possible. Definitions, if needed, should be provided in the question itself, because respondents often do not use rollovers or links for additional information when answering questions (Peytchev et al., 2010).

We suspect that applying the same guidelines to labeler prompts and instructions would improve the quality of the labels collected. (Of course, nothing can be done about ambiguous terms or high reading levels in the observations.) We are not aware of any research into these issues. We note that Figure 1 contains rollovers or links for additional information on three of the questions on the right side.

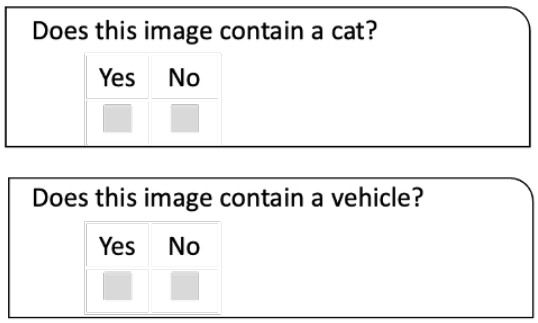
Multiple Labels Often ML researchers want to collect multiple labels about an observation from one labeler: for example, whether an image contains a cat, a dog, a person, or a vehicle. We can ask labelers to provide all labels at once, as in Figure 2(a), or we can ask one or more labelers to provide each label separately, as in Figure 2(b).

The choice between these two approaches echos the choice in surveys between *select-all-that-apply* questions (Figure 3(a)) and a series of *yes/no* questions (Figure 3(b)). As predicted by the survey response model, the *yes/no* format collects better data because it encourages respondents to process each option separately (Smyth, 2006; Pew Research Center, 2019). The *select-all-that-apply* approach is vulnerable to satisficing: respondents pick the first one or two reasonable options and fail to think deeply or even look at later options (Galesic et al., 2008). However, the *yes/no* format can encourage acquiescence (Smyth, 2006).

This finding also holds for labeling. In an experiment that involved labeling tweets as containing hate speech or offensive language, Kern et al. (2023) randomly assigned labelers to different versions of the labeling instrument. Condition A was similar to Figure 2(a) and Condition B was similar to Figure 2(b). Splitting the collection across two screens (Condition B) led to higher rates of hate speech and offensive language annotation. Models trained on Condition B data also performed better than those trained on Condition A data across several metrics (Kern et al., 2023). This result is a clear example of how findings in the survey literature translate to the labeling task and improve the quality of training data.

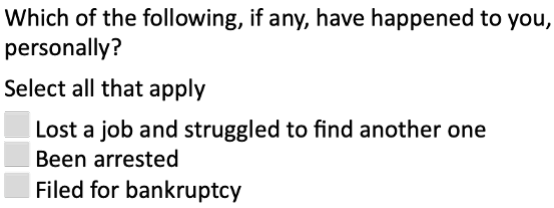


(a) Collect all labels on one screen

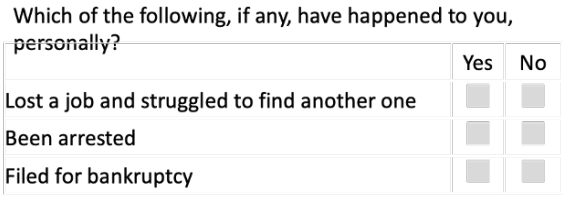


(b) Collect labels on separate screens

Figure 2. Collecting multiple labels on one screen (first panel) or multiple (second panel); adapted from (Kern et al., 2023)



(a) *Select-all-that-apply* question



(b) Series of *yes/no* questions

Figure 3. Survey question in *select all* (first panel) and *yes/no* (second panel) formats, adapted from Pew Research Center (2019)

Order Effects Theories about context effects suggest that the order in which instances or observations are presented influences the labels assigned. If a contrast effect is present, a very hateful social media post would make later posts seem less hateful than they otherwise would (for preliminary evidence of this phenomenon, see Beck et al., 2022). An order effect could also arise if labelers change their behavior over time. As they gain experience, they might become more accurate and conscientious, as suggested by Lee et al. (2022). Alternatively, they might become bored or fatigued and engage in more satisficing, acquiescence, or "don't know" nonresponding over time, as suggested by the survey literature (Kraut et al., 1975; Galesic & Bosnjak, 2009).

Mathur et al. (2017) detected order effects in two benchmark NLP data sets. When coding tweets as hate speech or offensive language, Beck et al. (2024) found a decreasing time trend: tweets that appeared later were less likely to be flagged. However, their experimental set up did not allow them to test hypotheses about the mechanisms causing the time trend.

Many research questions thus remain open, such as when contrast and assimilation effects appear and which tasks and labelers are most impacted by order effects. Order effects may also have implications for active learning (AL) and similar labeling approaches. In AL, an algorithm determines which observations to label next to maximize the marginal information gain for the model (Monarch, 2021). However, active learning considers only the model's needs, not the labelers'. If context effects are large, the algorithm should also account for contrast and assimilation effects when deciding which observations to label. We are not aware of any research that has jointly accounted for the needs of the model for diverse training data and the impact of observation order on annotators. An approach similar to active learning exists in surveys (Zhang et al., 2020), but is not widely adopted due in part to concerns about order effects.

Don't Know Option The inclusion of "don't know" or "no opinion" responses in surveys has been debated for years. Some researchers believe these options offer respondents an easy way to satisfice: rather than thinking about the issue and forming an opinion, respondents can simply choose the "don't know" option. Others believe that having no opinion on a given topic is a valid response and that forcing respondents to provide an opinion when they don't have one reduces data quality (Schuman & Presser, 1996).

Many labeling tasks do not include a "skip" or "don't know" option: labelers must provide a label even when they are not certain. (The instrument in Figure 1 is an exception.) When a recent experiment provided a "don't know" option to half of the labelers, fewer than three percent chose it, and the overall distribution of the labels was not impacted (Beck et al., 2022). Another recent study in NLP collected "uncertain" flags from labelers for a relation extraction task across several text genres (Bassignana & Plank, 2022). Labelers were more likely to choose "uncertain" when coding text in some genres, and the model struggled with prediction in those genres as well. These preliminary results suggest that giving labelers the option to indicate uncertainty or lack of knowledge can provide helpful information and does not encourage satisficing in labeling.

Pre-labeling Pre-labeling involves displaying a suggested label, bounding box, or similar and asking the labeler if it is correct. If the labeler indicates the label is not correct, they are asked to provide the correct label. Pre-labeling is more efficient than labeling without suggested labels (Lingren et al., 2013; Skeppstedt et al., 2016; South et al., 2014). However, labelers may become too trusting of the suggestions and fail to correct errors (Dietvorst et al., 2015; Logg, 2017; Berzak et al., 2016), a phenomenon called anchoring bias or automation bias (Mosier & Skitka, 1999).

In the survey field, we find that providing a pre-filled response that respondents or interviewers should update leads to underreporting of errors of both omission and commission. For example, when respondents are reminded of their answer in a previous survey wave, they tend to report that the answer still applies rather than providing an updated response (Jäckle & Eckman, 2019).

Previous literature on labeling has explored anchoring bias (Lingren et al., 2013; Skeppstedt et al., 2016; South et al., 2014) but has not leveraged social science to find the factors that make the effect weaker or stronger. The social science literature suggests several hypotheses about the mechanisms behind anchoring bias, such as incentives (Cialdini, 2009), belief in authority (Asch, 2016; Cialdini, 2009), or reliance on heuristics (Cialdini, 2009; Norman, 2007; Kahneman, 2011). Testing these theories experimentally would help data collectors design tasks that capture the efficiency of pre-labeling with lower risk of anchoring bias.

Overreliance on Examples Examples can introduce a similar bias. Survey questions often give examples of the things respondents should consider when they formulate their responses. Examples improve response accuracy when they remind respondents to include items they might otherwise leave out, because they have forgotten or were unsure whether to include them. However, when the examples include only common items, respondents tend to leave out less common items (Tourangeau et al., 2014).

Examples are also often included in labeling instructions or annotation guidelines. As in surveys, labelers at times rely too heavily on these examples as they label. This *instruction bias* can lead to overestimation of model performance

(Parmar et al., 2023). Again, a better understanding of the mechanisms behind this behavior, guided by social science theories, could inform efforts to reduce it.

3.3. Mitigation Measures and Future Research

The survey methods literature suggests several approaches to minimize the effects discussed in Section 3.2.

Randomization of Observations To address order effects, label collectors can randomize the order of observation shown to labelers. This approach does not eliminate order effect but it ensures that no one ordering impacts all annotators in the same way. Random ordering is incompatible with active learning techniques, however.

Instrument Testing Many surveys spend weeks or even months drafting, testing, and revising questions and response options to arrive at language that is understood similarly by most members of the population, a process called cognitive interviewing (Willis, 2004). They then launch the survey with a small group of respondents to assess response rates, don't know rates, and response times. Such testing could improve the instructions and prompts given to labelers.

Retain Paradata Many surveys capture process data, called paradata, during the survey, such as the time spent on each screen, the device used, even mouse movements (Kreuter, 2013; Horwitz et al., 2017). Paradata can help identify satisficing respondents and low quality data (Kreuter, 2013) and may do the same in label collection. However, collecting such data may raise additional privacy and ethics concerns (see Couper & Singer, 2012; Kunz et al., 2020; Henninger et al., 2023).

Feedback to Labelers Label collection instruments could experiment with prompts to encourage labelers not to engage in satisficing. Respondents who pick many "don't know" answers or repeatedly choose the same response option could receive reminders about the importance of the task. Those who click through screens quickly could receive prompts to slow down and read carefully. In surveys, feedback on speeding successfully slowed respondents and did not lead to early terminations of the survey (Conrad et al., 2017).

Test Observations In surveys, instructed response items, such as "Choose yellow below" can help identify respondents who speed or provide low quality responses (Gummer et al., 2018; Berinsky et al., 2024). We have not seen these questions used in labeling tasks. However, some tasks embed observations with known labels to try to catch annotators who do not understand the task. In NLP tasks, it is common to qualify only workers who pass an initial quiz or perform well on inserted test observations (Nangia et al., 2021). These test observations could also catch annotators who satisfice (see Nie et al., 2020, for an application).

We recommend future research to more comprehensively test these and related approaches.

4. Need for Diversity

Large-scale annotation tasks, such as the reCAPTCHA tests, may collect labels from a broad spectrum of the population. However, the crowdworkers used in many label collection tasks are members of large crowdworker panels such as Appen, Upwork, Scale AI, Prolific, or MTurk and do not reflect the U.S. or world population. Smart et al. (2024) note that labelers tend to be from the Global South, while the models they help train benefit the educated population in the Global North. The workers who labeled data to fine-tune InstructGPT were 22% Filipino and 22% Bangladeshi (Ouyang et al., 2022, Appendix B3). MTurk members are younger, lower income and less likely to live in the South than the U.S. population (Berinsky et al., 2012).

The unique characteristics of the labelers lead us to worry that the data they provide may not represent the views of the population that will use or be affected by the models.[2] Issues of representativeness are enormously important to surveys, which explicitly aim to make statements about the entire population. For this reason, the survey methods literature has much to contribute on this topic.

4.1. Selection Bias

Selection bias occurs when those involved in providing data have different characteristics than the population. In surveys, selection bias[3] arises when the propensity to take part in a survey is correlated with the characteristics measured in a survey. Let us return to the example in Section 2, a survey to estimate the proportion of U.S. adults without health insurance. If we distributed the survey invitations in doctors' waiting rooms, the sample proportion would overestimate the population proportion. Those in waiting rooms are more likely to receive the survey invitation and are also more likely to have health insurance. The propensity to take part is correlated with what the survey measures, leading to selection bias in the estimate of health insurance coverage.

[2] We acknowledge that this section glosses over what we mean by "population." Is it the population that regularly uses the models? The population impacted by the models? We leave this important discussion to later work.

[3] In the survey literature, the preferred terms is nonresponse bias. We use the term selection bias here because it is more general and more suited to the labeling task.

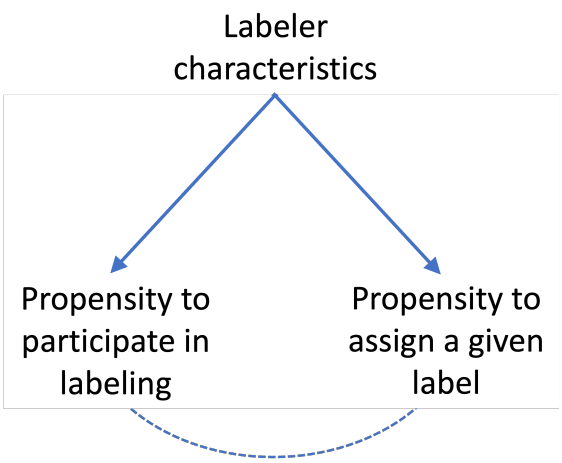


Figure 4. Labeler characteristics induce correlation between propensities (adapted from Groves, 2006)

Just as in surveys, selection bias arises in training data labels if the propensity to engage in the labeling task is correlated with the propensity to assign a given label. We expect that this correlation is non-zero for many tasks, because labeler characteristics likely influence both propensities, as shown in Figure 4.

On the left side of Figure 4, labelers' characteristics influence their decision to engage in labeling work. As noted above, labelers tend to have different attributes than the general population (Gray & Suri, 2019; Smart et al., 2024). On the right side of Figure 4, labelers' characteristics influence the labels they assign. The literature provides evidence for this association as well: labeler age and education level impact whether they perceive comments on Wikipedia entries as attacks (Al Kuwatly et al., 2020); conservative labelers are less likely to flag anti-Black language (Sap et al., 2022); labelers in the U.S. are more likely to see a bird in ambiguous images than those in India (Parrish et al., 2024). We hypothesize that such effects are more widespread than the literature suggests, because many studies do not collect annotator characteristics and thus cannot detect their impact on the labels (Kirk et al., 2023).

If the characteristics that influence one side of Figure 4 are the same as, or correlate with, the characteristics that influence the other side of the figure, the two propensities at the bottom will be correlated, leading to selection bias. Consider a labeler who is a frequent biker, annoyed by cars that park in the bike lanes in their city. They may be more likely to agree to label a data set of potential bike lane violations (left side) and also more likely to see violations where others do not (right side).

Although the goal of label data collection is not to make population estimates, selection bias is nevertheless a risk to training data and model development. In the early days of machine learning, when developers trained model to recognize written numbers or tell cats from dogs, perhaps labeler characteristics mattered less. The association on the right side of Figure 4 may be weak or absent with more objective tasks (though see Aroyo & Welty, 2015, for counterarguments).

However, the labeling tasks still performed by humans today often involve more difficult and more opinion-based work. People may legitimately disagree about whether a given statement is toxic or offensive, for example. Reinforcement learning with human feedback (RLHF) in particular may be more exposed to selection bias. The labels collected for reinforcement learning, like those in Figure 1, are inherently opinion-based. As discussed in Section 2, opinion questions in surveys are more susceptible to context effects than factual questions. We suspect that opinion labels are also more impacted by labeler characteristics than are more objective labels. RLHF aligns models to the judgments of labelers. If selection bias is present in the data, those judgments do not match the interests of the public. For example, if those who participate in labeling are also less likely to judge text as toxic, then the model trained on their data will also see less toxicity.

We see evidence of the impact of selection bias on models in two studies that trained models on different sets of labelers. Both showed that models make different predictions when trained on labels from, for example, female versus male labelers, or Asian versus white labelers (Al Kuwatly et al., 2020; Perikleous et al., 2022).

4.2. Mitigation Measures and Future Research

To combat selection bias in labeling tasks, we need to break the correlation between the propensity to assign a given label and the propensity to participate in labeling. We consider three methods.

Left Side We could try to remove the correlation between labeler characteristics and the propensity to participate in labeling (the left side of Figure 4) by diversifying the labeler pool, collecting data from labelers with different motivations and characteristics. This approach is central to surveys: we solicit responses from a random sample of the population through appeals to public service, tokens of appreciation, and multiple reminders (see, for example, Groves & Couper, 1998). If we can collect responses from a random sample of population members, selection bias is not a problem. Unfortunately, it is not clear that this approach will work with label collection: many people are not interested in labeling data for AI models. Surveys also find it increasingly difficult to recruit representative samples of respondents (de Leeuw et al., 2018; Williams & Brick, 2018).

Right Side Another way to reduce selection bias is to break the correlation on the right side of Figure 4, which would mean removing (or, more reasonably, reducing) the influence of labeler characteristics on the labels that they assign. More diverse examples in the instructions, use of test observations, feedback to labelers, and training in implicit bias might help labelers label more uniformly. (The literature on coding in qualitative studies takes a similar approach (see, for example, Hak & Bernts, 1996).) Even if these interventions do work to reduce the impact of labelers' characteristics on the labels they provide, however, they are expensive and do not scale well.

Interestingly, reducing selection bias by removing the correlation on the right side of the figure is not of interest in surveys, which aim to capture the diversity of respondents' behavior, opinions, and judgments. Recent research in NLP has similarly found that capturing the diversity of labels across labelers can improve models (Basile et al., 2021; Sap et al., 2022). Aroyo & Welty (2015) and Plank (2022) have argued that such human label variation is in fact *information* and can improve model performance and trustworthiness.

Weighting A third method to address selection bias is statistical adjustment. If we condition on the labeler characteristics in Figure 4, we can remove (or reduce) the induced correlation between the two propensities. Like the right side approach, this method involves embracing labeler subjectivity and uses weights to get the balance of those characteristics right.

The survey literature contains many statistical methods to match the characteristics of the respondents to the population and thus reduce selection bias (Bethlehem et al., 2011, for example). For example, surveys in many countries collect more responses from women than men; we use weights to ensure that the contribution of women's and men's responses on the final estimate matches their shares in the population. Future work could test the usefulness of these weighting approaches for improving machine learning models.[4]

However, these statistical adjustments can work only if we capture the labeler characteristics that drive the relationships in Figure 4. Thus we need a better understanding of what motivates people to work as labelers and what types of tasks are vulnerable to selection bias, which points to another role for social science to play in improving label quality.

[4]The tendency for some crowdworkers to give false answers to demographic questions to protect their privacy, reported by Huang et al. (2023), will complicate any weighting approach. Misreporting of demographics also causes problems in surveys (Pew Research Center, 2024).

5. Transparency in Label Collection

The discussion above suggests that labels are sensitive to how studies design the labeling task and recruit labelers, in ways often not recognized in the AI/ML literature. For this reason, we call for more transparency and documentation in how labels are collected when new data sets or models are released.

The survey industry in the U.S. has embraced transparency in recent years. The American Association for Public Opinion Research launched the Transparency Initiative in 2014. Member firms agree to disclose details about how survey data were collected, such as question and response option wording and order, respondent recruitment protocols, and weighting adjustments.[5] Polling companies that are members of the Transparency Initiative outperform those who are not (Silver, 2023), suggesting that a firm's willingness to disclose its data collection methods is a proxy for the quality of its estimates. The U.S. federal statistical agencies recently commissioned an expert report on transparency, in an effort to increase trust in federal data (National Academies of Sciences, Engineering, and Medicine, 2022).

Several researchers have similarly called for transparency when releasing benchmark data sets or models (Bender & Friedman, 2018; Mitchell et al., 2019; Hutchinson et al., 2020; Gebru et al., 2021; Chmielinski et al., 2022). We join these calls and recommend releasing the labeling instructions or guidelines including examples and test questions, the wording of the prompts, information about the labelers, and whether social scientists or domain experts were involved in labeling or consulted on the labeling process. Prabhakaran et al. (2021); Geiger et al. (2020); Ulmer et al. (2022); Baan et al. (2024) have also called for better documentation of the label collection process. We commend Nie et al. (2020); Ouyang et al. (2022); Glaese et al. (2022); Bai et al. (2022); Stiennon et al. (2022) as particularly good examples of transparency in label collection, with some including screenshots of the label collection instrument.

Without detailed documentation of data collection methods, researchers will not be able to test many of the hypotheses given above, such as those about wording, task order, "don't know" options, and the impact of labeler characteristics. We also suspect that lack of documentation explains difficulties replicating benchmark data sets (Recht et al., 2019).

6. Outlook

Collecting data from labelers is more difficult than the AI/ML literature has recognized. Ambiguous and opinion tasks are particularly challenging, because labelers' re-

[5]See https://aapor.org/wp-content/uploads/2022/11/TI-Attachment-C.pdf for details.

sponses can be shaped by wording, order, and other context effects as well as by the characteristics of the labeler. We believe that the type of human feedback needed to align future models will resemble opinion collection in surveys. For this reason, greater cooperation and knowledge sharing between the AI/ML and survey methods fields will be crucial to ensuring that the next generation of even more powerful models is human-centric, trustworthy, and fair.

Although some hope to replace human labelers with models-as-labelers (see, for example, Pangakis et al., 2023; Gilardi et al., 2023; Törnberg, 2023), the data collection challenges described in this paper will remain. Because models are trained on data collected from humans, as labelers, they can display many of the response biases described in this paper (Dominguez-Olmedo et al., 2023; Tjuatja et al., 2023). In addition, models trained on data labeled by models exhibit unusual behavior, called model collapse or model autophagy disorder (Alemohammad et al., 2023; Gerstgrasser et al., 2024; Peterson, 2024). Models trained on data labeled by models may also exhibit lower performance than those trained on data labeled by humans (see, for example, Plaza-del Arco et al., 2024). For the foreseeable future, we will still want humans to create the labels that teach models how to be accurate, fair and safe. Thus, the most important labeling must still be done by humans and is exactly the type of data that is most challenging to collect.

We applaud the growing interest in data-centric AI, which focuses on improving models by improving the data they rely on. We recommend that researchers developing AI models be as careful with how their data are collected as they are with their models. We also suggest reaching out to applied statistics departments and social science research groups which often focus on the science of data and may be willing to do the data work.

As is often the case when working across disciplines, as this paper does, language and cultural differences complicate information sharing. The survey methodology literature tends to focus on detecting and measuring bias and variance and attributing it to a cause (interviewers, question wording, survey mode). AI researchers, on the other hand, often understandably focus on solutions, in the form of algorithms distributed through python packages or github repos. These cultural differences between the fields can leave researchers unable to present at each other's conferences or even have productive conversations. Nevertheless, we hope that this position paper has made the case that theory-driven social science can help AI researchers develop insights and tools that improve the quality of their data and their models.[6] (And solution-oriented thinking could also make survey data better.) Perhaps future joint workshops could help the two fields share insights.

[6]We acknowledge that practitioners in AI and ML may be more aware of findings from survey research than the literature suggests.

Impact Statement

This position paper uses results from the field of survey methodology to derive hypotheses about the drivers of data quality in training data collection. We propose several areas for future research and provide concrete ideas to improve the quality of labels collected to train, fine-tune, reinforce, and evaluate AI and ML models. By improving the labels, these ideas will have positive effects on model performance and alignment.

Acknowledgments

In writing this paper, we benefited from discussions with Jacob Beck and Bolei Ma of LMU and Rob Chew of RTI International. Fiona Draxler provided valuable comments on an early draft. We would also like to thank the anonymous reviewers for their feedback. BP is supported by the European Research Council (ERC) grant agreements No. 101043235. This research is funded in part by the Bavarian Research Institute for Digital Transformation (bidt) project KLIMA-MEMES (BP) and GREEN DIA (FK), and the German Federal Ministry of Education and Research (BMBF) project KODAQS (16DKZ2019C). The authors are responsible for the content of this publication.

References

Al Kuwatly, H., Wich, M., and Groh, G. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 184–190, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.21. URL https://aclanthology.org/2020.alw-1.21.

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., and Baraniuk, R. G. Self-consuming generative models go mad, 2023.

Aroyo, L. and Welty, C. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36 (1):15–24, 2015. doi: 10.1609/aimag.v36i1.2564. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2564.

Asch, S. E. Effects of group pressure upon the modification and distortion of judgments. In *Organizational influence processes*, pp. 295–303. Routledge, 2016.

Baan, J., Fernandez, R., Plank, B., and Aziz, W. Interpreting predictive probabilities: Model confidence or human

label variation? In *Association for Computational Linguistics: EACL 2024*, Malta, March 2024. Association for Computational Linguistics.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., Uma, A., et al. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pp. 15–21. Association for Computational Linguistics, 2021.

Bassignana, E. and Plank, B. CrossRE: A cross-domain dataset for relation extraction. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3592–3604, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.263. URL https://aclanthology.org/2022.findings-emnlp.263.

Bates, N., Martin, E., DeMaio, T., and Puente, M. Questionnaire Effects on Measurement of Race and Spanish Origin. *Journal of Official Statistics*, 11(4):433–459, 1995.

Beck, J., Eckman, S., Chew, R., and Kreuter, F. Improving labeling through social science insights: Results and research agenda. In Chen, J. Y. C., Fragomeni, G., Degen, H., and Ntoa, S. (eds.), *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, pp. 245–261, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-21707-4.

Beck, J., Eckman, S., Ma, B., Chew, R., and Kreuter, F. Order effects in annotation tasks: Further evidence of annotation sensitivity. In Vázquez, R., Celikkanat, H., Ulmer, D., Tiedemann, J., Swayamdipta, S., Aziz, W., Plank, B., Baan, J., and de Marneffe, M.-C. (eds.), *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pp. 81–86, St Julians, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.uncertainlp-1.8.

Bender, E. M. and Friedman, B. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041.

Berinsky, A. J., Huber, G. A., and Lenz, G. S. Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis*, 20(3): 351–368, 2012. doi: 10.1093/pan/mpr057.

Berinsky, A. J., Frydman, A., Margolis, M. F., Sances, M. W., and Valerio, D. C. Measuring Attentiveness in Self-Administered Surveys. *Public Opinion Quarterly*, 88(1):214–241, March 2024. ISSN 1537-5331. doi: 10.1093/poq/nfae004. URL https://doi.org/10.1093/poq/nfae004. _eprint: https://academic.oup.com/poq/article-pdf/88/1/214/57226568/nfae004.pdf.

Berzak, Y., Huang, Y., Barbu, A., Korhonen, A., and Katz, B. Anchoring and agreement in syntactic annotations. In Su, J., Duh, K., and Carreras, X. (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2215–2224, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1239. URL https://aclanthology.org/D16-1239.

Bethlehem, J., Cobben, F., and Schouten, B. *Handbook of Nonresponse in Household Surveys*. Wiley, January 2011. ISBN 9780470891056. doi: 10.1002/9780470891056. URL http://dx.doi.org/10.1002/9780470891056.

Bless, H. and Schwarz, N. Mental construal and the emergence of assimilation and contrast effects: The inclusion/exclusion model. *Advances in Experimental Social Psychology*, 42:319–373, 12 2010. doi: 10.1016/S0065-2601(10)42006-7.

Brosnan, K., Babakhani, N., and Dolnicar, S. "i know what you're going to ask me" why respondents don't read survey questions. *International Journal of Market Research*, 61(4):366–379, 2019. doi: 10.1177/1470785318821025. URL https://doi.org/10.1177/1470785318821025.

Chmielinski, K. S., Newman, S., Taylor, M., Joseph, J., Thomas, K., Yurkofsky, J., and Qiu, Y. C. The dataset nutrition label (2nd gen): Leveraging context to mitigate harms in artificial intelligence, 2022.

Cialdini, R. B. *Influence : science and practice*. Pearson Education, Boston, 2009. ISBN 0205609996 9780205609994 9780205663781 0205663788. URL http://www.amazon.co.uk/Influence-Practice-Robert-B-Cialdini/dp/0205663788.

Clinton, J., Agiesta, J., Brenan, M., Burge, C., Connelly, M., Edwards-Levy, A., Fraga, B., Guskin, E., Hillygus, D. S., Jackson, C., Jones, J., Keeter, S., Khanna, K., Lapinski, J., Saad, L., Shaw, D., Smith, A., Wilson, D., and Wlezien, C. Task force on 2020 pre-election polling: An evaluation of the 2020 general election polls, 2021.

Conrad, F., Tourangeau, R., Couper, M., and Zhang, C. Reducing speeding in web surveys by providing immediate feedback. *Survey Research Methods*, Vol 11:No 1 (2017), 2017. doi: 10.18148/SRM/2017.V11I1.6304. URL https://ojs.ub.uni-konstanz.de/srm/article/view/6304.

Couper, M. P. and Singer, E. Informed consent for web paradata use. *Survey Research Methods*, 7(1):57–67, Dec. 2012. doi: 10.18148/srm/2013.v7i1.5138. URL https://ojs.ub.uni-konstanz.de/srm/article/view/5138.

Cowan, C., Murphy, L., and Wiener, J. Effects of Supplemental Questions on Victimization Estimates from the National Crime Survey. In *Proceedings of the Survey Research Methods Section*, pp. 277–282. American Statistical Association, 1978. URL http://www.asasrms.org/Proceedings/papers/1978_055.pdf.

de Leeuw, E., Hox, J., and Luiten, A. International nonresponse trends across countries and years: An analysis of 36 years of labour force survey data. *Survey Insights, Methods from the Field (SMIF)*, 2018.

Dietvorst, B. J., Simmons, J. P., and Massey, C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology General*, 2015. doi: 10.1037/xge0000033.

Dillman, D. A., Smyth, J. D., and Christian, L. M. *Internet, phone, mail, and mixed-mode surveys*. John Wiley & Sons, Nashville, TN, 4 edition, August 2014.

Dominguez-Olmedo, R., Hardt, M., and Mendler-Dünner, C. Questioning the survey responses of large language models, 2023.

Eckman, S., Kreuter, F., Kirchner, A., Jäckle, A., Tourangeau, R., and Presser, S. Assessing the mechanisms of misreporting to filter questions in surveys. *Public Opinion Quarterly*, 78(3):721–733, 2014. doi: 10.1093/poq/nfu030. URL https://doi.org/10.1093/poq/nfu030.

Fiske, S. T. and Taylor, S. E. *Social Cognition*. McGraw Hill Higher Education, Maidenhead, England, 2 edition, April 1991.

Galesic, M. and Bosnjak, M. Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey. *Public Opinion Quarterly*, 73(2): 349–360, 05 2009. ISSN 0033-362X. doi: 10.1093/poq/nfp031. URL https://doi.org/10.1093/poq/nfp031.

Galesic, M., Tourangeau, R., Couper, M. P., and Conrad, F. G. Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly*, 72(5):892–913, December 2008. ISSN 1537-5331. doi: 10.1093/poq/nfn059. URL http://dx.doi.org/10.1093/poq/nfn059.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, November 2021. ISSN 1557-7317. doi: 10.1145/3458723. URL http://dx.doi.org/10.1145/3458723.

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., and Huang, J. 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. ACM, January 2020. doi: 10.1145/3351095.3372862. URL http://dx.doi.org/10.1145/3351095.3372862.

Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Sleight, H., Hughes, J., Korbak, T., Agrawal, R., Pai, D., Gromov, A., Roberts, D. A., Yang, D., Donoho, D. L., and Koyejo, S. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data, 2024.

Gibson, C. O., Shapiro, G. M., Murphy, L. R., and Stanko, G. J. Interaction of Survey Questions as it Relates to Interviewer-Respondent Bias. In *Proceedings of the Survey Research Methods Section*, pp. 251–256. American Statistical Association, 1978. URL http://ww2.amstat.org/sections/srms/Proceedings/papers/1978_050.pdf.

Gilardi, F., Alizadeh, M., and Kubli, M. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), July 2023. ISSN 1091-6490. doi: 10.1073/pnas.2305016120. URL http://dx.doi.org/10.1073/pnas.2305016120.

Glaese, A., McAleese, N., Trebacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P., Campbell-Gillingham, L., Uesato, J., Huang, P.-S., Comanescu, R., Yang, F., See, A., Dathathri, S., Greig, R., Chen, C., Fritz, D., Elias, J. S.,

Green, R., Mokrá, S., Fernando, N., Wu, B., Foley, R., Young, S., Gabriel, I., Isaac, W., Mellor, J., Hassabis, D., Kavukcuoglu, K., Hendricks, L. A., and Irving, G. Improving alignment of dialogue agents via targeted human judgements, 2022.

Gray, M. and Suri, S. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt, 2019. ISBN 978-1-328-56624-9. URL https://books.google.com/books?id=8AmXDwAAQBAJ.

Groves, R. M. Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70(5): 646–675, 2006. ISSN 0033-362X. doi: 10.1093/poq/nfl033. URL http://dx.doi.org/10.1093/poq/nfl033.

Groves, R. M. and Couper, M. P. *Nonresponse in Household Interview Surveys*. Wiley, April 1998. ISBN 9781118490082. doi: 10.1002/9781118490082. URL http://dx.doi.org/10.1002/9781118490082.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. *Survey Methodology*. Wiley Series in Survey Methodology. Wiley-Blackwell, Hoboken, NJ, 2 edition, June 2009.

Gummer, T., Roßmann, J., and Silber, H. Using instructed response items as attention checks in web surveys: Properties and implementation. *Sociological Methods & Research*, 2018. doi: 10.1177/0049124118769083.

Hak, T. and Bernts, T. Coder training: Theoretical training or practical socialization? *Qualitative Sociology*, 19(2): 235–257, June 1996. ISSN 1573-7837. doi: 10.1007/bf02393420. URL http://dx.doi.org/10.1007/BF02393420.

Henninger, F., Kieslich, P. J., Fernández-Fontelo, A., Greven, S., and Kreuter, F. Privacy Attitudes toward Mouse-Tracking Paradata Collection. *Public Opinion Quarterly*, 87(S1):602–618, 08 2023. ISSN 0033-362X. doi: 10.1093/poq/nfad034. URL https://doi.org/10.1093/poq/nfad034.

Horwitz, R., Kreuter, F., and Conrad, F. Using mouse movements to predict web survey response difficulty. *Social Science Computer Review*, 35(3):388–405, 2017. ISSN 0894-4393. doi: 10.1177/0894439315626360. URL https://doi.org/10.1177/0894439315626360.

Huang, F. L. and Cornell, D. G. Question order affects the measurement of bullying victimization among middle school students. *Educational and Psychological Measurement*, 2015. doi: 10.1177/0013164415622664. URL http://epm.sagepub.com/content/early/2015/12/16/0013164415622664.abstract.

Huang, O., Fleisig, E., and Klein, D. Incorporating worker perspectives into MTurk annotation practices for NLP. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1010–1028, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.64. URL https://aclanthology.org/2023.emnlp-main.64.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., and Mitchell, M. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure, 2020.

Irving, G. and Askell, A. Ai safety needs social scientists. *Distill*, 4(2), February 2019. ISSN 2476-0757. doi: 10.23915/distill.00014. URL http://dx.doi.org/10.23915/distill.00014.

Jäckle, A. and Eckman, S. Is that still the same? has that changed? on the accuracy of measuring change with dependent interviewing. *Journal of Survey Statistics and Methodology*, 8(4):706–725, July 2019. ISSN 2325-0992. doi: 10.1093/jssam/smz021. URL http://dx.doi.org/10.1093/jssam/smz021.

Kahneman, D. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, 2011. ISBN 9780374275631 0374275637.

Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., McGeeney, K., Miringoff, L., Olson, K., Rivers, D., Sadd, L., Witt, E., and Wlezien, C. An evaluation of 2016 election polls in the united states, 2017.

Kern, C., Eckman, S., Beck, J., Chew, R., Ma, B., and Kreuter, F. Annotation sensitivity: Training data collection methods affect model performance. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14874–14886, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-emnlp.992.

Kim, Y., Dykema, J., Stevenson, J., Black, P., and Moberg, D. P. Straightlining: Overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Soc. Sci. Comput. Rev.*, 37(2):214–233, April 2019.

Kirk, H. R., Vidgen, B., Röttger, P., and Hale, S. A. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback, 2023.

Knowles, E. S. and Condon, C. Why people say "yes": a dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, 77:379–386, 1999. doi: 10.1037/0022-3514.77.2.379.

Kraut, A. I., Wolfson, A. D., and Rothenberg, A. Some effects of position on opinion survey items. *J. Appl. Psychol.*, 60(6):774–776, December 1975.

Kreuter, F. *Improving Surveys with Paradata*. John Wiley & Sons, Inc., Hoboken, New Jersey, September 2013.

Kreuter, F., McCulloch, S., Presser, S., and Tourangeau, R. The Effects of Asking Filter Questions in Interleafed versus Grouped Format. *Sociological Methods and Research*, 40(88):88–104, 2011.

Krosnick, J. A. Response Strategies for Coping With the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, 5(5):213–236, 1991. URL https://web.stanford.edu/dept/communication/faculty/krosnick/docs/1991/1991%20Satisficing.pdf.

Krosnick, J. A., Narayan, S., and Smith, W. R. Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, 1996(70):29–44, March 1996. doi: 10.1002/ev.1033. URL https://doi.org/10.1002/ev.1033.

Kunz, T., Landesvatter, C., and Gummer, T. Informed consent for paradata use in web surveys. *International Journal of Market Research*, 62(4):396–408, 2020. doi: 10.1177/1470785320931669. URL https://doi.org/10.1177/1470785320931669.

Lee, J.-U., Klie, J.-C., and Gurevych, I. Annotation curricula to implicitly train non-expert annotators. *Computational Linguistics*, 48(2):343–373, June 2022. doi: 10.1162/coli_a_00436. URL https://aclanthology.org/2022.cl-2.4.

Lingren, T., Deleger, L., Molnar, K., Zhai, H., Meinzen-Derr, J., Kaiser, M., Stoutenborough, L., Li, Q., and Solti, I. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. *Journal of the American Medical Informatics Association*, 21(3):406–413, 09 2013. ISSN 1067-5027. doi: 10.1136/amiajnl-2013-001837. URL https://doi.org/10.1136/amiajnl-2013-001837.

Logg, J. M. Theory of machine: When do people rely on algorithms?", 2017. URL http://nrs.harvard.edu/urn-3:HUL.InstRepos:31677474.

Mathur, N., Baldwin, T., and Cohn, T. Sequence effects in crowdsourced annotations. In Palmer, M., Hwa, R., and Riedel, S. (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2860–2865, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1306. URL https://aclanthology.org/D17-1306.

Mazumder, M., Banbury, C., Yao, X., Karlaš, B., Rojas, W. G., Diamos, S., Diamos, G., He, L., Parrish, A., Kirk, H. R., Quaye, J., Rastogi, C., Kiela, D., Jurado, D., Kanter, D., Mosquera, R., Ciro, J., Aroyo, L., Acun, B., Chen, L., Raje, M. S., Bartolo, M., Eyuboglu, S., Ghorbani, A., Goodman, E., Inel, O., Kane, T., Kirkpatrick, C. R., Kuo, T.-S., Mueller, J., Thrush, T., Vanschoren, J., Warren, M., Williams, A., Yeung, S., Ardalani, N., Paritosh, P., Bat-Leah, L., Zhang, C., Zou, J., Wu, C.-J., Coleman, C., Ng, A., Mattson, P., and Reddi, V. J. Dataperf: Benchmarks for data-centric ai development, 2023.

McFarland, S. G. Effects of question order on survey responses. *Public Opinion Quarterly*, 45(2):208–215, 01 1981. ISSN 0033-362X. doi: 10.1086/268651. URL https://doi.org/10.1086/268651.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, January 2019. doi: 10.1145/3287560.3287596. URL http://dx.doi.org/10.1145/3287560.3287596.

Monarch, R. M. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Manning, 2021. URL https://www.manning.com/books/human-in-the-loop-machine-learning.

Mosier, K. L. and Skitka, L. J. Automation use and automation bias. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 43(3):344–348, 1999. doi: 10.1177/154193129904300346. URL https://doi.org/10.1177/154193129904300346.

Nangia, N., Sugawara, S., Trivedi, H., Warstadt, A., Vania, C., and Bowman, S. R. What ingredients make for an effective crowdsourcing protocol for difficult NLU data collection tasks? In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1221–1235, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.98. URL https://aclanthology.org/2021.acl-long.98.

National Academies of Sciences, Engineering, and Medicine. Transparency in statistical information for the national center for science and engineering statistics and all federal statistical agencies, 2022. URL https://doi.org/10.17226/26360.

Nie, Y., Zhou, X., and Bansal, M. What can we learn from collective human opinions on natural language inference data?, 2020.

Norman, D. *Emotional design*. Basic Books, London, England, March 2007.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, 2022.

Pangakis, N., Wolken, S., and Fasching, N. Automated annotation with generative ai requires validation, 2023.

Parmar, M., Mishra, S., Geva, M., and Baral, C. Don't blame the annotator: Bias already starts in the annotation instructions. In Vlachos, A. and Augenstein, I. (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1779–1789, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.130. URL https://aclanthology.org/2023.eacl-main.130.

Parrish, A., Hao, S., Laszlo, S., and Aroyo, L. Is a picture of a bird a bird? a mixed-methods approach to understanding diverse human perspectives and ambiguity in machine vision models. In Abercrombie, G., Basile, V., Bernadi, D., Dudy, S., Frenda, S., Havens, L., and Tonelli, S. (eds.), *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pp. 1–18, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.nlperspectives-1.1.

Perikleous, P., Kafkalias, A., Theodosiou, Z., Barlas, P., Christoforou, E., Otterbacher, J., Demartini, G., and Lanitis, A. How does the crowd impact the model? a tool for raising awareness of social bias in crowdsourced training data. In *Proceedings of the 31st ACM International Conference on Information; Knowledge Management*, CIKM '22. ACM, October 2022. doi: 10.1145/3511808.3557178. URL http://dx.doi.org/10.1145/3511808.3557178.

Peterson, A. J. Ai and the problem of knowledge collapse, 2024.

Pew Research Center. When online survey respondents only select some that apply, 2019. URL https://www.pewresearch.org/methods/2019/05/09/when-online-survey-respondents-only-select-some-that-apply/.

Pew Research Center. Online opt-in polls can produce misleading results, especially for young people and hispanic adults, 2024. URL https://www.pewresearch.org/short-reads/2024/03/05/online-opt-in-polls-can-produce-misleading-results-especially-for-young-people-and-hispanic-adults/.

Peytchev, A., Conrad, F. G., Couper, M. P., and Tourangeau, R. Increasing respondents' use of definitions in web surveys. *Journal of Official Statistics*, 26:633–650, 2010.

Plank, B. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10671–10682, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.731. URL https://aclanthology.org/2022.emnlp-main.731.

Plaza-del Arco, F. M., Nozza, D., and Hovy, D. Wisdom of instruction-tuned language model crowds. exploring model label variation. In Abercrombie, G., Basile, V., Bernadi, D., Dudy, S., Frenda, S., Havens, L., and Tonelli, S. (eds.), *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pp. 19–30, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.nlperspectives-1.2.

Prabhakaran, V., Mostafazadeh Davani, A., and Diaz, M. On releasing annotator-level labels and information in datasets. In Bonial, C. and Xue, N. (eds.), *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pp. 133–138, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.law-1.14. URL https://aclanthology.org/2021.law-1.14.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet?, 2019.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. M. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for

Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445518. URL https://doi.org/10.1145/3411764.3445518.

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., and Smith, N. A. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5884–5906, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.431. URL https://aclanthology.org/2022.naacl-main.431.

Schuman, H. and Presser, S. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage, 1996.

Schwarz, N. Cognitive aspects of survey methodology. *Applied Cognitive Psychology*, 21(2):277–287, February 2007. ISSN 1099-0720. doi: 10.1002/acp.1340. URL http://dx.doi.org/10.1002/acp.1340.

Silver, N. How our pollster ratings work, March 2023. URL https://fivethirtyeight.com/methodology/how-our-pollster-ratings-work/.

Skeppstedt, M., Paradis, C., and Kerren, A. Pal, a tool for pre-annotation and active learning. *Journal for Language Technology and Computational Linguistics*, 2016. doi: 10.21248/jlcl.31.2016.203.

Smart, A., Wang, D., Monk, E., Díaz, M., Kasirzadeh, A., Liemt, E. V., and Schmer-Galunder, S. Discipline and label: A weird genealogy and social theory of data annotation, 2024.

Smyth, J. D. Comparing check-all and forced-choice question formats in web surveys. *Public Opinion Quarterly*, 70(1):66–77, March 2006. ISSN 1537-5331. doi: 10.1093/poq/nfj007. URL http://dx.doi.org/10.1093/poq/nfj007.

South, B. R., Mowery, D., Suo, Y., Leng, J., Óscar Ferrández, Meystre, S. M., and Chapman, W. W. Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. *Journal of Biomedical Informatics*, 50:162–172, 2014. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2014.05.002. URL https://www.sciencedirect.com/science/article/pii/S1532046414001191. Special Issue on Informatics Methods in Medical Privacy.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. Learning to summarize from human feedback, 2022.

Strack, F. "order effects" in survey research: Activation and information functions of preceding questions. In Schwarz, N. and Sudman, S. (eds.), *Context Effects in Social and Psychological Research*, pp. 23–34, New York, NY, 1992. Springer New York. ISBN 978-1-4612-2848-6. doi: 10.1007/978-1-4612-2848-6_3. URL https://doi.org/10.1007/978-1-4612-2848-6_3.

Sturgis, P., Baker, N., Callegaro, M., Fisher, S., Green, J., Jennings, W., Kuha, J., Lauderdale, B., and Smith, P. Report of the inquiry into the 2015 british general election opinion polls, 2016.

Tjuatja, L., Chen, V., Wu, S. T., Talwalkar, A., and Neubig, G. Do llms exhibit human-like response biases? a case study in survey design, 2023.

Todorov, A. The accessibility and applicability of knowledge: Predicting context effects in national surveys. *Public Opinion Quarterly*, 64(4):429–451, 2000. doi: 10.1086/318639. URL http://poq.oxfordjournals.org/content/64/4/429.abstract.

Tourangeau, R. The survey response process from a cognitive viewpoint. *Quality Assurance in Education*, 26(2):169–181, April 2018. ISSN 0968-4883. doi: 10.1108/qae-06-2017-0034. URL http://dx.doi.org/10.1108/QAE-06-2017-0034.

Tourangeau, R., Rips, L. J., and Rasinski, K. A. *The psychology of survey response*. Cambridge University Press and Cambridge Univ. Press, 10. print edition, 2000. ISBN 978-0-521-57246-0.

Tourangeau, R., Conrad, F. G., Arens, Z., Fricker, S., Lee, S., and Smith, E. Everyday concepts and classification errors: Judgments of disability and residence. *Journal of Official Statistics*, 22:385–418, 2006.

Tourangeau, R., Kreuter, F., and Eckman, S. Motivated underreporting in screening interviews. *Public Opinion Quarterly*, 76(3):453–469, August 2012. doi: 10.1093/poq/nfs033. URL https://doi.org/10.1093/poq/nfs033.

Tourangeau, R., Conrad, F., Couper, M., and Ye, C. The effects of providing examples in survey questions. *Public Opinion Quarterly*, 78:100–125, 03 2014. doi: 10.1093/poq/nft083.

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

Törnberg, P. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning, 2023.

Ulmer, D., Bassignana, E., Müller-Eberstein, M., Varab, D., Zhang, M., van der Goot, R., Hardmeier, C., and Plank, B. Experimental standards for deep learning in natural language processing research. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2673–2692, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.196. URL https://aclanthology.org/2022.findings-emnlp.196.

Williams, D. and Brick, J. M. Trends in u.s. face-to-face household survey nonresponse and level of effort. *J. Surv. Stat. Methodol.*, 6(2):186–211, June 2018.

Willis, G. B. *Cognitive interviewing*. Sage Publications, Christchurch, New Zealand, September 2004.

Zaller, J. and Feldman, S. A simple theory of the survey response: Answering questions versus revealing preferences. *American Journal of Political Science*, 36(3):579, August 1992. ISSN 0092-5853. doi: 10.2307/2111583. URL http://dx.doi.org/10.2307/2111583.

Zhang, C., Taylor, S., Cobb, C., and Sekhon, J. Active matrix factorization for surveys. *Annals of Applied Statistics*, 14:1182–1206, 09 2020. doi: 10.1214/20-AOAS1322.