Jacob Beck

Improving Annotation Quality: Empirical Insights into Bias, Human-AI Collaboration, and Workflow Design

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 24.06.2025

Jacob Beck

Improving Annotation Quality: Empirical Insights into Bias, Human-AI Collaboration, and Workflow Design

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 24.06.2025

Erste Berichterstatterin: Prof. Dr. Frauke Kreuter
Zweite Berichterstatterin: Prof. Dr. Barbara Plank
Dritter Berichterstatter: Prof. Dr. Trent Buskirk

Tag der Disputation: 06.10.2025

Acknowledgments

I would like to thank ...

- *... my "doctoral parents", Prof. Frauke Kreuter and Dr. Stephanie Eckman:*
  Thank you, Frauke, for the freedom you gave me, the support you provided, and especially for the trust you placed in me. These conditions created an environment where I could grow personally and academically.
  Thank you, Stephanie, for your invaluable guidance at every stage and for always appreciating my viewpoints and perspectives.
- *... Prof. Barbara Plank and Prof. Trent Buskirk for their time and effort as referees of the dissertation, and Prof. Göran Kauermann and Prof. Joseph Sakshaug for their engagement as examiners at the dissertation defense.*
- *... my writing group, Olga and Markus, and my annotation research group, Rob, Christoph, and Bolei. Meeting you on a regular basis helped me stay on track, get the occasional needed kick, and most importantly, keep my sanity.*
- *... everyone in the FK2RG group, the LMU Statistics Department, university staff, my external project partners, all involved in the wonderful DSSGx Munich programs, and the "Statistical Football" group.*
- *... the many people beyond the academic world who have taught me something – whether about a subject, a skill, or life itself.*
- *... my family for their confidence they had in me, and above all, Jana for supporting, challenging, and loving me every day.*

Summary

High-quality annotated datasets are essential for training machine learning (ML) models. Annotation means assigning a label (such as a category, sentiment score, or classification) to an instance, for example to a piece of text, an image, or a PDF file. Even as training algorithms continue to improve, a model's real-world performance remains limited by the quality of the training data. While there are many approaches for processing training data, relatively little attention within the ML field has been devoted to annotation quality and the development of best practices for data collection. This thesis contributes to the field through empirical assessments of annotation bias and its implications for training data quality. It further proposes and evaluates strategies to mitigate such biases and enhance annotation outcomes. In addition, it explores the role of large language models (LLMs) in annotation workflows by experimentally assessing their use in fully automated and human-assisted hybrid annotation pipelines.

The introductory part outlines the research questions and motivates the overall contributions. As part of this, the background chapter provides a review of the literature on factors influencing annotation quality, organized along two main dimensions: Annotator-related factors encompass individual-level traits and behaviors that may be correlated with annotation behavior. Annotation data collection strategies refer to all design-related decisions made when setting up a task, such as the selection of examples provided in the instructions, task length, or payment. In addition, challenges and opportunities of automating annotation are discussed.

Annotation is a structured task that follows standardized procedures for data collection, typically involving a stimulus and fixed response options, much like data collection in fields such as survey methodology and social psychology. In the first and second study, we investigate whether well-known sources of bias identified in these fields also apply to annotation tasks. The first study presents experimental results from a large sample of annotators. We analyze task structure and demographic effects in a hate speech sentiment annotation task, systematically varying the screen design to measure its effect on the resulting labels. In addition, we collect demographic characteristics, task perception metrics, and paradata to assess their relationship with label assignment. Most notably, annotation behavior was significantly influenced by whether classification tasks appeared on a single screen or were split across two, as well as by the annotator's first language. The second study extends this project by examining whether annotation behavior changes over the course of the task. It estimates how the likelihood of assigning a label evolves with the number of previously completed annotations. As the task progressed, labeling a statement as hateful or offensive became significantly less likely, though the effect was small in magnitude. Together, these studies show that annotations are sensitive to both who performs them and how the task is structured.

The third and fourth study explore the potential of real-time, low-cost automated annotations generated by LLMs and their interaction with human annotators. In the third study, we conduct a cost-benefit analysis comparing different types of human and automated annotators in a satellite image annotation task. It includes initial attempts to combine human and LLM-generated annotations. We observe strong potential for cost reduction and quality retention, with less need for expert annotators – especially when leveraging the LLM's self-reported uncertainty. The fourth study builds on this study by documenting a pipeline for generating and curating a gold-standard validation dataset for CO_2 emission values extracted from PDF documents. It demonstrates a

feasible approach to integrating automated components to reduce the workload of human domain experts. Even in this highly specialized task, combining LLM annotations with non-expert adjudication can substantially reduce reliance on domain experts.

The fifth study investigates the risks and implications of increasing automation in annotation workflows, particularly pre-annotations generated by artificial intelligence (AI). We simulate an AI-assisted scenario by presenting annotators with pre-annotations framed as AI-generated, to examine cognitive bias during adjudication. Notably, those who reported greater skepticism toward AI were more accurate in adjudicating the pre-annotations. Additionally, we observe that annotators are less likely to correct pre-annotations when flagging an error requires providing a corrected value.

Across its five contributions, this dissertation advances the field of annotation data collection methods by identifying bias in human, automated, and hybrid annotation setups. It proposes and evaluates multiple solutions and offers guidance for both research and practical annotation tasks. A consistent focus is placed on integrating insights and theories from various academic disciplines to benefit from a broad range of existing findings.

Zusammenfassung

Qualitativ hochwertige annotierte Datensätze sind für das Training von Modellen des maschinellen Lernens (ML) unerlässlich. Annotation bedeutet, dass einer Instanz (z.B. einem Text, einem Bild oder einer PDF-Datei) ein Label (wie eine Kategorie, Bewertung oder Klassifikation) zugewiesen wird. Auch wenn Trainingsalgorithmen sich stetig verbessern, begrenzt die Qualität der Trainingsdaten in der Praxis die Leistung eines Modells. Während es viele Ansätze zur verbesserten Verarbeitung von Trainingsdaten gibt, konzentriert sich die Forschung im Bereich ML weniger auf die Qualität der Annotationen und die Entwicklung von Best Practices für die Datenerhebung. Die vorgelegte Dissertation trägt zu diesem Forschungsfeld bei, indem sie die Verzerrung von Annotationen und deren Auswirkungen auf die Qualität der Trainingsdaten empirisch untersucht. Darüber hinaus werden Strategien vorgeschlagen und evaluiert, um solche Verzerrungen abzuschwächen und damit die Qualität der annotierten Daten zu verbessern. Darüber hinaus wird die Rolle von "Large Language Models" (LLMs) in Annotations-Workflows untersucht, indem ihre Verwendung in voll automatisierten und von Menschen assistierten hybriden Annotations-Pipelines experimentell untersucht wird.

Der einleitende Abschnitt skizziert die übergeordneten Forschungsfragen und begründet deren Relevanz. In diesem Zusammenhang wird ein Literaturüberblick über die Faktoren gegeben, die die Qualität von annotierten Daten beeinflussen. Dieser Überblick ist nach zwei zentralen Dimensionen gegliedert: Annotator-bezogene Faktoren umfassen individuelle Eigenschaften und Verhaltensweisen, die mit dem Annotationsverhalten korrelieren können. Annotationsstrategien beziehen sich auf alle gestaltungsbezogenen Entscheidungen, die bei der Erstellung einer Annotierungsaufgabe getroffen werden, wie z.B. die Auswahl der in den Anweisungen enthaltenen Beispiele, die Länge der Aufgabe oder die Bezahlung. Außerdem werden Herausforderungen und Chancen der Automatisierung von Annotationen erörtert.

Die Annotation von Daten ist eine strukturierte Aufgabe, die standardisierten Verfahren zur Datenerhebung folgt und in der Regel einen Stimulus und festgelegte Antwortoptionen umfasst, ähnlich wie bei der Datenerhebung in Bereichen wie der Umfrageforschung und der Sozialpsychologie. In der ersten und zweiten Studie untersuchen wir, ob bekannte Quellen der Verzerrung, die in diesen Bereichen identifiziert wurden, auch für Annotationsaufgaben gelten. Die erste Studie präsentiert experimentelle Ergebnisse aus einer großen Stichprobe von Annotatoren. Wir analysieren die Aufgabenstruktur und demografische Effekte in einer Annotationsaufgabe von Hassrede, wobei wir systematisch das Bildschirmdesign variieren, um die Auswirkungen auf die resultierenden Annotationen zu messen. Darüber hinaus werden demografische Merkmale, Metriken zur Aufgabenwahrnehmung und Paradaten gesammelt, um ihre Beziehung zur Label-Zuweisung zu bewerten. Das Annotationsverhalten wurde vor allem davon signifikant beeinflusst, ob die Frame oder in zwei aufeinanderfolgenden Schritten präsentiert wurden, sowie von der Muttersprache des Annotators. Die zweite Studie erweitert dieses Projekt, indem sie untersucht, ob sich das Annotationsverhalten im Verlauf einer Aufgabe ändert. Sie schätzt ab, wie sich die Wahrscheinlichkeit, ein Label zu vergeben, mit der Anzahl der zuvor abgeschlossenen Annotationen entwickelt. Mit dem Fortschreiten der Aufgabe wurde es deutlich unwahrscheinlicher, dass eine Aussage als hasserfüllt oder beleidigend eingestuft wurde, auch wenn die Effektgröße gering ausfiel. Zusammengenommen zeigen diese beiden Studien, dass Annotationen sowohl davon abhängen, wer sie vornimmt, als auch davon, wie die Aufgabe strukturiert ist.

Die dritte und vierte Studie untersuchen das Potenzial von automatisierten Annotationen, die von LLMs nahezu in Echtzeit generiert werden, und deren Kombination mit menschlichen Annotatoren. In der dritten Studie führen wir eine Kosten-Nutzen-Analyse durch, in der wir verschiedene Arten von menschlichen und automatisierten Annotatoren bei einer Aufgabe zur Annotation von Satellitenbildern vergleichen. Sie umfasst erste Versuche, menschliche und LLM-generierte Annotationen zu kombinieren. Wir stellen fest, dass es ein großes Potenzial zur Kostenreduzierung unter Bewahrung der Datenqualität gibt. Der Bedarf an Experten-Annotatoren kann vor allem dann verringert werden, wenn die vom LLM selbst angegebene Unsicherheit genutzt wird. Die vierte Studie baut auf dieser Studie auf, indem sie eine Pipeline zur Erhebung und Aufbereitung eines Gold-Standard-Validierungsdatensatzes für CO_2-Emissionswerte aus PDF-Dokumenten dokumentiert. Sie demonstriert einen praktikablen Ansatz zur Integration automatisierter Komponenten, um den Arbeitsaufwand menschlicher Fachexperten in der Annotation von Daten zu reduzieren. Selbst bei dieser hochspezialisierten Aufgabe kann die Kombination von LLM-Annotationen mit der Beurteilung durch Nicht-Experten die Abhängigkeit von Domänenexperten erheblich reduzieren.

Die fünfte Studie untersucht die Risiken und Auswirkungen der zunehmenden Automatisierung von Annotations-Workflows, insbesondere von Vor-Annotationen, die durch künstliche Intelligenz (KI) generiert werden. Wir simulieren ein KI-gestütztes Szenario, indem wir den menschlichen Annotatoren Vor-Annotationen vorlegen, die als KI-generiert dargestellt sind, um die kognitive Verzerrung bei der Beurteilung dieser zu untersuchen. Bemerkenswert ist, dass diejenigen, die eine größere Skepsis gegenüber künstlicher Intelligenz angaben, die Vor-Annotationen akkurater bewerteten. Wir stellen außerdem fest, dass die Annotatoren weniger geneigt sind, Vor-Annotationen zu korrigieren, wenn das Markieren eines Fehlers die Angabe eines korrigierten Wertes erfordert.

Mit ihren fünf Beiträgen bringt diese Dissertation den Bereich der Methoden zur Erhebung von annotierten Daten voran, indem sie Verzerrungen in menschlichen, automatisierten und hybriden Annotationskonzepten aufzeigt. Sie schlägt mehrere Lösungen vor, bewertet diese und bietet Orientierungshilfen sowohl für die Forschung als auch für praktische Annotationsaufgaben. Ein zentraler Schwerpunkt liegt auf der Integration von Erkenntnissen und Theorien aus verschiedenen akademischen Disziplinen, um von der Breite bestehender Forschungsergebnisse zu profitieren.

Contents

Part I.

Introduction and Background

1. Introduction

1.1. Motivation

At its core, any artificial intelligence (AI) model relies on two fundamental components: the data and the algorithm that learns patterns from it. While algorithms and model architectures continue to advance, the true value of an AI system lies in its effectiveness beyond controlled benchmarks, in real-world settings where data is more complex, noisy, and unpredictable. This effectiveness depends heavily on the quality of the training data used to teach these models – specifically, the quality of how this data has been annotated with labels (such as categories, ratings, or classifications). The quality depends on how accurate and informative these labels are, and how well the dataset reflects the distribution of real-world data samples. Data quality is often the more decisive factor, outweighing the marginal gains achieved through further algorithmic improvements. Although this dependency on data quality is widely acknowledged, those responsible for designing annotation pipelines often face a large number of complex decisions, ranging from task design to annotator selection, with limited standardized guidance available when collecting and annotating real-world data. The central problem is twofold: the sources of bias in annotation pipelines remain underexplored, and as a result, actionable and sound strategies for mitigating bias are limited. These challenges are further complicated by the high degree of diversity within the field of annotation. It spans a wide range of data modalities, from text and images to audio and video. Importantly, annotation tasks vary in their degree of subjectivity, with some requiring objective labeling and others involving more subjective judgment. This distinction matters because these two types pursue fundamentally different goals. Objective annotation tasks aim to detect the true label. In contrast, subjective (or "perspectivist" (Fleisig et al., 2024; Frenda et al., 2024)) annotation tasks intend to model the meaningful signal within annotator disagreement (Plank, 2022). Rather than treating disagreement as noise, subjective annotation acknowledges that it stems not just from annotator error, but also from subjective differences in how annotators see the same instance. Moreover, annotation efforts differ based on constraints such as the required domain expertise of the annotators, their availability, and the broader context in which the annotation takes place.

There are still significant gaps in our understanding of bias in training data annotation and in the methods used to mitigate it. The introduction and rapid advancement of large language models (LLMs) add new layers of complexity, raising new questions rather than resolving existing ones. Progress in this area is happening so quickly that research aimed at identifying biases and related problems often struggles to keep pace. Yet, addressing these questions is crucial, particularly as LLMs increasingly make the case for either replacing human annotators or supporting them in collaborative annotation workflows (Aguda et al., 2024; Goel et al., 2023; Li et al., 2023; Li, 2024), and rely on annotated data themselves for their training and continual improvement.

One underexplored perspective that can help us better understand annotation, identify sources of bias, and develop robust practices is to place the data – and the processes that generate it

– at the center of analysis. This approach, which one might call a "science of data" approach, rejects the notion that data naturally appears in the wild as a neutral or incidental byproduct or that its generation and preparation are merely procedural steps (Ang et al., 2013). A "science of data" view on annotation brings a strong interdisciplinary foundation, drawing on the extensive methodological knowledge that other fields have developed over decades. Annotation is essentially a task designed to guide humans through a standardized process, typically involving a stimulus and a fixed set of response options. In this sense, it closely resembles a (web) survey. At the same time, annotation is often a form of human-computer interaction (HCI), and automated annotation can be seen as a specific instance of algorithmic decision-making (ADM). Therefore, beyond the social and behavioral sciences, related fields such as HCI, ADM, and survey methodology offer valuable frameworks (Biemer, 2010; Jones-Jang and Park, 2022; Lyberg et al., 2012; Rastogi et al., 2022) for understanding training data quality in machine learning (ML). These disciplines also contribute practical methods and tools for identifying, measuring, and addressing challenges in the data generation process (Biemer et al., 2017; Dimara et al., 2019; Liu et al., 2024; Rieger et al., 2021; Wyer, 2010).

This thesis takes an interdisciplinary approach to annotation bias and its implications for training data quality. Drawing on multiple academic perspectives, it contributes to the field of data annotation by identifying sources of bias and examining strategies for mitigation. The empirical work brings the field closer to understanding and improving annotation practices, with particular attention to recent developments in automation and hybrid annotation setups, where human annotators collaborate with automated systems.

1.2. Outline

The thesis features five articles, listed in Table 1.1. In the next chapter of Part I., I will outline existing research and motivate the overarching research questions, building on and extending the literature synthesis presented in Beck (2023). The empirical work is divided into three parts.
Part II. Human Annotation Sensitivity examines whether theoretical perspectives on bias, well established in the social sciences and particularly in survey methodology, are transferable to the process of annotating data. Specifically, we investigate how task structure, instance order, and annotator demographics influence the resulting annotations (Article 1 and Article 2).
Part III. Automation in Annotation explores two implementations of automated LLM-based annotation in real-world pipelines. These implementations are evaluated in terms of their quality, measured by how closely they approximate expert annotations, as well as their potential and risks in hybrid annotation setups (Article 3 and Article 4). The studies cover two different data modalities: satellite imagery of industrial land and PDF documents from company reports.
Building on these findings, Part IV. Bias in Human-AI Collaborative Annotation shifts focus to investigating new sources of bias that may arise when such automation is embedded into annotation workflows. This question is explored through a user study assessing how human annotators interact with, and are influenced by, AI-generated pre-annotations (Article 5).

Table 1.1.: Overview of Contributing Articles

Article	Title	Authors	Publication Status
A1	Improving Labeling Through Social Science Insights: Results and Research Agenda	Jacob Beck, Stephanie Eckman, Rob Chew, Frauke Kreuter	Published in *Proceedings of HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence* (2022)
A2	Order Effects in Annotation Tasks: Further Evidence of Annotation Sensitivity	Jacob Beck, Stephanie Eckman, Bolei Ma, Rob Chew, Frauke Kreuter	Published in *Proceedings of the 1st Workshop on Uncertainty-Aware NLP* (2024)
A3	Toward Integrating ChatGPT into Satellite Image Annotation Workflows	Jacob Beck, Lukas Malte Kemeter, Konrad Dürrbeck, Mohamed Hesham Ibrahim Abdalla, Frauke Kreuter	Published in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2025)
A4	Addressing Data Gaps in Sustainability Reporting	Jacob Beck, Anna Steinberg, Andreas Dimmelmeier, Laia Domenech Burin, Emily Kormanyos, Maurice Fehr, Malte Schierholz	Published in *Nature Scientific Data*
A5	Bias in the Loop: How Humans Respond to AI-Generated Pre-Annotations	Jacob Beck, Stephanie Eckman, Frauke Kreuter	Revised version under review at *Harvard Data Science Review*

2. Background

2.1. Introduction

Traditionally, ML models are trained on data instances with one or more labels. Assigning these labels is a task known as *annotation*.[1] A model learns to predict the label of an unlabeled instance and is then typically evaluated by assessing how well it predicts labels on a hold-out portion of the training data. However, learning from and reproducing erroneous data causes the model to learn patterns that do not reflect the real-world task. For example, an annotator may misclassify the industrial activity of land in satellite images by mistaking construction vehicles for industrial activity, leading the model to learn incorrect associations. Annotation errors of this kind can substantially limit the value of a model's practical implementation, because models can learn from incorrectly assigned labels. When the same type of error occurs in both training and testing data, performance metrics cannot detect the problem. In practice, systematic annotation errors fundamentally limit real-world performance – particularly when incorrect or overly simplistic labels are used, or when the training data distribution doesn't match deployment conditions. Consequently, the design of the annotation process should minimize errors and ensure that the dataset used for model generation accurately reflects the target distribution.

But how can this be achieved? Annotating data is by no means a uniform process. Tasks differ substantially depending on the data modality (for example text vs. image), the level of domain expertise required, or the degree of subjectivity. Whether consciously or not, those designing annotation tasks make a multitude of decisions throughout the development of the annotation pipeline. These efforts are often constrained by available resources or contextual factors. In some cases, the pool of suitable and available annotators is limited, budgets are tight, time pressure is high, or data-sharing restrictions prevent collaboration beyond institutional boundaries, narrowing the range of feasible choices. Nevertheless, the number of open design decisions remains large, and those who collect annotated data are frequently left without clear guidance. In the absence of such guidance, unrecognized sources of bias can easily enter the data collection process and affect downstream outcomes. Addressing these sources of annotation bias requires identifying and measuring them in order to develop effective mitigation strategies.

In this chapter[2], I outline previous research and approaches to the collection of annotated training data. I highlight studies from diverse academic disciplines in which parameters that affect the quality of annotated data are identified, estimated, discussed, or accounted for. These studies collectively stress the necessity for thoughtful consideration of the annotation process among both researchers and practitioners. The background chapter is structured around two key dimensions of data annotation and its quality confounders: the set of *annotators* and the *strategy* of data

[1] Throughout this thesis, the term *labeling* is used interchangeably with *annotation*.

[2] This background chapter draws extensively on my published literature review "Quality aspects of annotated data: A research synthesis" (Beck, 2023), which has been adapted and extended for this thesis.

collection. The first section features studies that examine the connection between annotator composition and behavior, and the resulting data. The second section outlines different strategies for constructing, implementing, and evaluating an annotated data collection process. It highlights the many decisions involved in data collection, most of which lack established best practices, or are only addressed in highly task- and data-specific ways. Both sections are further divided into subsections that cover specific aspects of data annotation.

Importantly, the growing role of automation has fundamentally influenced the field of annotation, integrating both the strategic and annotator-related dimensions. Automated annotation, such as through LLMs, and hybrid annotation pipelines, in which humans collaborate with AI systems, offer compelling advantages in terms of cost, efficiency, and scalability. However, these approaches also introduce novel threats and uncertainties to data quality that must be critically examined and addressed. The third section of the background chapter therefore explores existing work on the automation of annotation processes.

The background chapter concludes by highlighting key insights from prior work and demonstrating how the research questions addressed in my featured articles are situated within the identified gaps.

2.2. Annotators

A common way to obtain training data for ML models is through human data annotation. Annotators differ in their backgrounds, attitudes, past experiences, worldviews, and other characteristics that shape how they interpret and annotate data. In principle, anyone can be an annotator. In practice, annotators most commonly include researchers, domain experts, company employees, student assistants, and crowdworkers.

The following section provides an overview of commonly encountered annotator profiles, followed by a focused discussion of two particularly relevant types: crowdworkers and domain experts. I then examine annotators from two perspectives: annotator characteristics and annotator behavior. Within each subsection, I highlight concepts that may affect the quality of annotated data.

2.2.1. Annotator Profile

Before addressing annotators' specific characteristics and behavioral patterns in general, it is crucial to acknowledge the diversity of annotator profiles. These profiles encompass different roles, such as crowdworkers, student assistants, researchers, and domain experts. Each profile offers unique contributions and faces specific challenges. Notably, some annotator profiles may be more susceptible to specific biases that I will discuss in this section. Among these diverse profiles, crowdworkers and domain experts represent two particularly significant ones for ML data annotation applications. Crowdworkers provide a large-scale workforce at low cost, are quickly available, and work across any domain. In contrast, domain experts are annotators from whom researchers expect the highest annotation data quality, though they often have limited availability, fewer exist, and require higher costs. Due to the specific characteristics of these two profiles, a large share of human annotation applications have previously employed either crowdworkers or experts as annotators. The following discussion will examine specific observations and mechanisms for these two groups.

Crowdworking platforms such as Amazon Mechanical Turk or Prolific increasingly organize the distribution of crowdworking labor (Belletti et al., 2021; Cefkin et al., 2014). These platforms serve as quick and efficient tools to split data annotation into microtasks and retrieve the required annotations through human crowdworkers. In typical settings, researchers or companies act as task requesters who set up tasks and provide the pool of crowdworkers with microtasks. Crowdworkers then choose to work on specific tasks in exchange for payment that the task requesters usually define in advance. Since researchers frequently employ crowdworkers as annotators, most existing research examining phenomena around annotation focuses on crowdsourced[3] annotations. Generally, crowdsourcing allows retrieval of annotations at high velocity with relatively little cost and effort. Wang et al. (2013) outline various approaches to engage crowdworkers, from gamified platforms to collective intelligence systems. However, many factors raise doubts about the inherent data quality of crowdsourced annotations. These factors include the (precarious) work standards, the incentive structure, and the commitment of annotators. Additionally, the susceptibility to unwanted bot annotations threatens data quality and replicability. To address these concerns, crowdworking platforms constantly work to improve data quality and increasingly provide relevant information such as annotator demographics or metadata (such as response times). Moreover, crowdsourced annotation increasingly benefits from well-designed interventions that enable bias mitigation and data quality improvement (Zhang et al., 2017).

When annotation tasks require specialized expertise, they are often assigned to professionals with relevant domain knowledge. This profile of annotators usually consists of domain experts whom researchers recruit for the annotation (for example, doctors involved in skin cancer classification or in-house experts working on proprietary data) or who annotate data for their own ML application. Here, the relation between the annotator and the application can play an important role in the expert's identification and resulting motivation for the task. While domain experts often provide high-quality annotations, their limited availability and high cost can lead to overreliance on a small number of individuals. Furthermore, assuming these annotations are always correct increases the risk of overlooking potential biases.

No ideal annotator profile exists. The choice of profile depends on task-specific requirements, resource availability, and contextual constraints. Moreover, researchers lack systematic comparisons between annotator profiles. This gap makes it difficult to develop clear guidelines for selecting the most appropriate profile for a given task.

2.2.2. Annotator Characteristics

Annotators typically represent a narrow and task-specific subset of individuals. They either self-select into data annotation (for example, on crowdsourcing platforms) or receive assignments to annotation tasks (such as doctors who must label medical documents). The socio-demographic composition of annotators on Amazon MTurk shows more balance than typical convenience samples, such as self-selected college students, but less alignment with the national population than high-quality internet panels or probability panels (Berinsky et al., 2012). Examining the demographic composition of annotator pools raises a fundamental question: does representativeness

[3] In this dissertation, crowdsourcing refers specifically to the practice of obtaining annotations from crowdworkers – individuals who complete small tasks for payment on online platforms, rather than the broader definition of sourcing ideas or content from a large group of people.

matter for annotation quality? In traditional web surveys, demographic representation is crucial because most surveys aim to draw inferences about specific populations. Annotation tasks, however, serve different purposes than surveys. Researchers typically do not collect annotated datasets to make population-level inferences, meaning that annotator pools do not necessarily need to constitute random population samples. Nevertheless, even when population representativeness is not the primary concern, annotator characteristics can still significantly influence the resulting datasets. Annotations may vary systematically based on who provides them, potentially introducing unwanted biases into the data. Previous studies have identified and analyzed various annotator characteristics that affect annotation quality. This subsection outlines how these characteristics impact annotated datasets.

Expertise Expertise or the individual's qualification for an annotation task plays an important role in annotation behavior and resulting data quality. While expertise is ideally a continuous variable, researchers and developers typically distinguish between laypersons (such as crowdworkers or student assistants) and domain experts (for example, radiologists for annotating X-ray images). This distinction matters from both a data quality and resource efficiency perspective. Layperson annotations are typically easier and cheaper to obtain. However, many assume that expert annotations provide higher quality results, and that certain tasks, such as X-ray image classification, cannot be reasonably completed by non-experts. While laypersons may need to build their reasoning from scratch, experts can rely on existing knowledge and beliefs (Heerkens et al., 2011), and this may affect both the speed and consistency of their decisions. Researchers would ideally assess the quality difference between expert and layperson annotations to determine which approach better serves the resulting model. While this assessment remains task-specific and difficult to quantify, it helps guide decisions. This decision also depends on whether objective domain experts exist for the task in question (such as hate speech detection) and whether the task actually requires domain expertise (for example, does classifying images of cats and dogs need a biologist?).

Research comparing the quality gap between expert and non-expert annotators continues to grow. For example, expert and layperson annotators achieved comparable agreement scores in an occupation-coding task, which suggests limited added value of expert input in that context (Maaz et al., 2009). Many studies show that researchers have developed methods that can be applied to improve the quality of non-expert annotations. These methods help non-experts achieve performance levels comparable to experts in various tasks. (Aroyo and Welty, 2015; Dumitrache et al., 2015; Heim et al., 2018; Nowak and Rüger, 2010). For example, they reframe complex classification tasks (like fish species identification) into simpler visual similarity judgments that become more accessible to non-experts (He et al., 2013; Wang and Vasconcelos, 2023; Yang et al., 2019).

First Language The annotator's first language could be an important demographic feature for judging the outcome of a language annotation task. Generally, this variable serves as a proxy for language proficiency, given the difficulty of measuring proficiency levels directly. If a task requires language understanding, proficiency level should logically be a key determinant of annotator aptitude. However, many annotation platforms and tasks do not restrict annotators based on language proficiency requirements. Crowdworkers who complete tasks in English live around the globe and often do not have to meet formal language prerequisites. In 2009, 36% of the

Amazon MTurk workforce lived in India (Ross et al., 2010). Since English is the first language to only 0.02% of India's population (Singh et al., 2022), most Indian MTurkers are likely not first language English speakers. In one study, 48% of the individuals who labeled English-language tweets for hate speech were Venezuelan residents (Founta et al., 2018). While learning English as a second/foreign language does not automatically mean insufficient language proficiency, complex, multilayered tasks like hate speech detection seem to require a very high degree of English understanding to create high-quality annotated data. Annotators should be able to grasp slang, irony, and sarcasm, in addition to cultural understanding (Bui et al., 2025). This pattern is evident in recent research: non-native English speakers labeled significantly fewer tweets as hateful compared to native English speakers, with the study limited to US residents only (Beck et al., 2022).

Further evidence shows that data quality differs based on annotator language background (Al Kuwatly et al., 2020). In this study, annotators examined whether comments on Wikipedia contained personal attacks. The researchers then grouped annotations by native and non-native English speakers and trained separate models on each dataset. Models trained on data from native English speakers proved significantly more sensitive than those trained on non-native speaker annotations.

Race/Ethnicity Few studies have systematically analyzed how annotators' racial or ethnic identities may influence annotated data quality, and the literature that does exist shows mixed findings. For example, Arhin et al. (2021) found that Black annotators deviated more frequently from the majority label in a toxic text classification task. Similarly, Larimore et al. (2021) observe significant differences in annotations between White and non-White annotators when assessing the racial sentiment of tweets.

Independent of annotator identity, models predict higher toxicity for statements in African American English (AAE) compared to non-AAE statements (Sap et al., 2019). In a subsequent experiment, instructing annotators to consider both the dialect and the racial or ethnic background of a statement's creator resulted in fewer toxic annotations. Various datasets contain racial and ethnic bias, ranging from hate speech detection – identified using topic modeling (Davidson and Bhattacharya, 2020) – to image captioning (Zhao et al., 2021). Raising awareness of these forms of bias in annotation and training data helps prevent models from picking up racist patterns and reinforce them when deployed. When researchers detect bias, they can apply post-hoc bias mitigation methods (Xia et al., 2020).

Gender No significant differences in model sensitivity and specificity were found for models trained on male and female annotators' data, respectively (Al Kuwatly et al., 2020). Consistent with this finding, annotated data sets did not meaningfully differ by gender across four different Natural Language Processing (NLP) tasks (Biester et al., 2022). In a study that used a previously annotated corpus of Wikipedia comments to train models that predicted the toxicity of statements, annotator gender led to small differences in the resulting training data. However, models trained on each group produced very similar results (Binns et al., 2017). In contrast, studies found clear gender differences in toxicity annotation (Excell and Al Moubayed, 2021), facial recognition tasks (Chen and Joo, 2021), offensive language and racism annotation (Sap et al., 2022) and sentiment analysis across four different annotation modalities (Ding et al., 2022). Regarding toxicity/hate

speech, women took a more negative stance towards the harm of hate speech and, on average, valued freedom of speech as less important than men did (Cowan and Khatchadourian, 2003).

Age Depending on the task, annotator age may influence annotation patterns. If model outputs vary by age and the annotator sample deviates from broader population distributions, analyzing annotations by age becomes essential to identify and mitigate potential distortions. Following this approach, Al Kuwatly et al. (2020) detected significant differences in both sensitivity and specificity between models trained on annotations that they grouped by age. An assessment of Amazon MTurk annotator characteristics revealed the age distribution of MTurkers to be significantly younger than that of other convenience samples and the US national population (Berinsky et al., 2012).

Political Orientation An annotator's political orientation can serve as a proxy for beliefs and values held by that individual, which could be especially important in subjective annotation tasks. Argument annotations in two political contexts (cloning and minimum wage) differed significantly by political leaning of annotators. The study measured the political orientation as self-reported categorization of conservative or liberal. These differences in annotations transferred downstream into algorithmic bias (Thorn Jakobsen et al., 2022). In addition, conservative annotators annotated AAE as toxic more frequently while simultaneously flagging fewer instances of racist language as toxic (Sap et al., 2022). More abstractly, stereotypes held by annotators correlated with their hate speech annotation behavior and the resulting classifier errors (Davani et al., 2023). In 2012, a sample of Amazon MTurkers leaned more democratic and more liberal compared to the US national distribution (Berinsky et al., 2012). While this distribution might have shifted in the past years, crowdworkers still likely comprise a demographically unbalanced sample, including with respect to political orientation. Such imbalances might affect the outcome of certain types of (more subjective) annotation tasks and the consequent models.

Conclusion When and how annotator characteristics affect the data generation process remains unclear. For some characteristics, empirical findings are inconclusive – for example, in the case of gender – while others, such as education, still lack sufficient research (Al Kuwatly et al., 2020). Additionally, less visible factors beyond standard demographics or assumed expertise, such as annotators' personal beliefs, can influence outcomes, even in tasks typically considered objective (Beck et al., 2025a). Despite these uncertainties, collecting and carefully monitoring differences in annotator characteristics is still important for understanding and mitigating potential biases. The relevance of demographic or individual traits can vary significantly depending on the annotation task. While having expert biologists label genome sequences rather than using a probability sample from the general population makes sense, models that should inherit societal beliefs and values can become distorted by heavily biased annotator samples.

2.2.3. Annotator Behavior

Annotation resembles surveys in that both present individuals with a stimulus (survey questions and annotation instances, respectively) and a set of fixed choices (response categories or label options). This structural similarity suggests that many well-studied conscious and subconscious

cognitive processes that influence survey responses may also play a role in annotation tasks. For example, first impressions can significantly anchor perceptions and prove resistant to change, even when people receive new information (Harris et al., 2023; Rabin and Schrag, 1999; Ybarra, 2005). This phenomenon relates closely to the well-established concept of confirmation bias, in which individuals tend to prioritize information that aligns with their preexisting beliefs while discounting or ignoring contradictory evidence (Nickerson, 1998; Oswald and Grosjean, 2004). In the context of data annotation, studies have examined some of these cognitive processes, such as anchoring or confirmation bias (Eickhoff, 2018; Hube et al., 2019). Other phenomena, such as speeding (annotating at an unreasonable velocity) or straightlining (repeatedly selecting the same label option regardless of the presented instance), are well-known in survey methodology but remain underexplored in annotation research (Schonlau and Toepoel, 2015; Zhang and Conrad, 2014). These mechanisms stem from general principles of human cognition and should, in principle, apply to all types of annotators, though not necessarily to the same extent. The following paragraphs illustrate three additional social psychological concepts that may influence data annotation behavior, with particular relevance for crowdworkers, who represent the most extensively studied annotator group.

Motivation Understanding what motivates individuals who participate in annotation tasks helps task requesters design annotation tasks in line with annotators' motivations and, if applicable, use additional motivating factors. Motivations can range from purely monetary incentives to intrinsic interest in the resulting model (for example, when annotating data for one's own research). A systematic analysis of interactions and conversations in a large Amazon MTurk forum, "Turker Nation", showed that monetary motivations were by far the most important motivating factor among crowdworkers (Martin et al., 2014). The enjoyability of a task impacted its popularity – for example, workers accepted slightly lower-paid tasks were accepted if they found them enjoyable – but the monetary aspect remained essential to the forum participants. Beyond a more positive perception of enjoyable tasks, annotators became more active when a task was framed as meaningful (Chandler and Kapelner, 2013). To convey such a sense of meaningfulness, the task told some annotators that their work would contribute to medical research, while it gave no context or informed them that their annotations would be discarded after the task. The perception of meaningfulness linked to increased participation rates, annotation volume, and data quality (Chandler and Kapelner, 2013).

While the importance of monetary motivations naturally characterizes crowdworking – which platforms specifically advertise as an easy way to earn money – task requesters should keep in mind that annotators do not necessarily want to create high-quality data or well-performing models. Survey methodologists have developed theories and practical approaches on how to collect and evaluate survey participation reasons that could apply to and benefit annotation tasks (Haensch et al., 2022; Keusch, 2015; Singer, 2011). However, crowdworkers may misreport their motivations for engaging in annotation work due to social desirability bias (Antin and Shaw, 2012). Social desirability bias likely affects response behavior when reporting motivations because of power asymmetries between crowdworker and task requester that arise from crowdworking being a crucial source of income for many crowdworkers (Martin et al., 2014; Miceli et al., 2022).

Dishonesty Dishonest behavior or misreporting can occur in surveys and in annotation tasks, driven by individual motivations and incentives. In crowdsourced annotation tasks in particular,

multiple reasons exist for dishonest behavior. People may submit incorrect information to meet the eligibility criteria for an annotation task or to reduce the perceived task burden. This phenomenon is called "motivated misreporting" (Eckman et al., 2014; Kreuter et al., 2011; Tourangeau et al., 2012). If such behaviors do not occur at random, the resulting training data becomes prone to bias, and overall data quality suffers from dishonest annotator behavior. Identifying task elements that encourage misreporting can help researchers and task designers create annotation tasks hat minimize the likelihood of such behavior.

Several studies have examined dishonest annotation behavior. For example, annotators willingly provided wrong answers for better payment, and fraudulent behavior decreased when they sensed being detected (Suri et al., 2011). Crowdworkers may misreport individual characteristics to be admitted to an annotation task (Chandler and Paolacci, 2017). In a prescreening survey for a task that required being a parent of an autistic child, respondents reported such parenthood approximately twice as often as in the same task where this criterion was not mandatory. Consistent with this finding, a similar experiment in the same study shows that at certain payment levels, respondents reported a different gender for study eligibility (Chandler and Paolacci, 2017).

Networking among Annotators When trying to understand annotating behavior and the self-selecting process of annotators, task designers need to consider networking and information exchange between annotators. Crowdworkers use online forums to exchange annotation strategies and information with others (Martin et al., 2014). Forum users share ways to earn money easier and faster, as well as which tasks are more enjoyable. The community generally condemns fraudulent behavior or cheating but not the use of loopholes within tasks or the exploitation of tasks with low payment – for instance, through reduced effort in the annotation process. Furthermore, annotators share insights about task requesters they consider good or bad (Martin et al., 2014). While not everyone participates actively in online forums and forum users represent a self-selected sample, the assumption of independence between observations (i.e., annotation responses) may not be valid. This concern grows stronger because crowdworking accounts may be shared by multiple individuals rather than being tied to a single user.

2.3. Data Collection Strategy

Every decision regarding the annotation collection strategy may impact the resulting set of annotations and the subsequently trained models. These decisions span the entire data collection process and range from considerations about required annotation sample sizes to task design or data evaluation approaches. Building on insights into annotator characteristics and behavior, the following section outlines four central decision areas for data annotation.

2.3.1. Task Design

Different task designs can lead annotators towards different annotation patterns (Pyatkin et al., 2023). When designing annotation tasks, many decisions must be made that may affect the resulting data quality. Although these decisions might seem minor, they often lack empirical grounding and instead reflect arbitrary choices. Without clearly understanding whether, and how, certain design features of annotation tasks affect annotation behavior, such choices affect

the resulting data and models (Kern et al., 2023). This subsection showcases a range of task design options and potential effects on data quality.

Label Options Determining which and how many label options to provide is not always straightforward or clear from the data or model. The level of annotation detail can range from simple binary classifications to continuous scales or open-ended class additions. The number of label options can vary depending on the desired degree of aggregation – that is, how broadly or narrowly categories are defined – that annotation achieves (Maaz et al., 2009) and the intended use of the resulting model. Label aggregation presents a clear trade-off between the information gained and the cognitive burden placed on annotators, as increasing the number of label options generally increases both (Kutlu et al., 2020). One potential adjustment to the label scale involves adding an option that allows annotators to express uncertainty, such as a "don't know" label. While this label option prevents forcing annotators to select an unsuitable label, it may also encourage them to avoid making a decision, using the option whenever they feel slight doubt. Empirically, the value of such an option remains unclear (Beck et al., 2022). Similarly, adding a residual category (for example "other" or "none of the above") to the label set can reduce misclassifications in cases where an exhaustive list of labels is not feasible and a catch-all category is required.

Rationale Asking annotators to provide rationale behind every annotation judgment can improve the quality of the resulting data and yield additional information (Kutlu et al., 2020). However, more experienced crowdworkers (who completed 20 or more tasks) were less likely to spend the additional time to provide the rationale, when requested but not mandatory (Kutlu et al., 2020). While asking for rationale does not appear feasible for a full-scale annotation process, requesting annotators' judgment rationale might help the data collection process at an earlier stage. A more feasible application could involve requesting rationales in a potential "pre-test" setting of an annotation task. Similar to conducting cognitive interviews (Beatty and Willis, 2007), where participants express their full thought process in preparation for experiments or surveys, studies could collect a smaller number of annotations with extensive rationale prior to the main data collection. This approach helps identify problematic annotation behavior, check for consistent interpretation of label categories, and assess the effectiveness of guidelines and examples. Insights from this phase can inform improvements to task design, much like in responsive survey design (Groves and Heeringa, 2006).

Guidelines Another component of an annotation task that can potentially bias or anchor the subsequent annotation process are the initial annotation guidelines or tutorials. Guidelines represent an essential, yet often resource-intensive, component of the annotation design process within a theoretical framework (Fort, 2016). These guidelines can exert significant leverage because many annotators read the same guidelines that task designers construct only once, and similar guidelines often apply across multiple annotation tasks. When constructing task tutorials, requesters must make important decisions such as the number and selection of examples while balancing the degree of leeway they give to annotators. Empirical results show that using annotation guidelines improves the quality of annotated data (Nédellec et al., 2006). Additionally, the way task designers formulate these instructions influences annotations and annotator bias (Thorn Jakobsen et al., 2022).

Order The order in which instances are presented to annotators matters for two key reasons. First, previously perceived information influences the perception of current content, which aligns with the theory of contrast and assimilation (Bless and Schwarz, 2010). For example, annotators labeled tweets as less hateful when they saw them after a more hateful tweet, compared to the same tweet that followed a less hateful one. This pattern provides initial evidence for a contrast effect (Beck et al., 2022). A contrast effect means that annotators perceive an instance as more dissimilar to the previously annotated instance(s). As a result, their judgments depend heavily on the already seen data (Bless and Schwarz, 2010).

The second ordering effect works independently of content but relates to exposure over time, suggesting that annotation behavior changes as a task progresses, for example due to fatigue or learning effects. Annotators became less likely to flag tweets as hateful or offensive as the annotation task went on (Beck et al., 2024).

Therefore, annotation items should be presented in random order to minimize bias that order effects introduce. While random ordering represents a best practice for many applications, it could potentially be problematic for non-random ordering approaches like Active Learning (AL). AL describes an ML approach where the model predicts which instance's annotation would currently provide the model with the greatest benefit in terms of model performance (Settles, 2009). This purposeful ordering (by the model) could foster unwanted order effects. However, AL generally offers multiple benefits, such as reducing annotation costs. Therefore, we must weigh the expected bias that non-random ordering introduces against the anticipated AL benefits (Zhang et al., 2022). Additionally, we need to assess the ordering of multiple different tasks empirically. If annotators must make two annotations for one instance (for example, the brightness and the resolution of an image), designers must decide on the task order. They can retrieve both annotations on one screen, have brightness annotations followed by resolution annotations, or annotate each image for brightness first, then immediately follow with that image's resolution annotation.

Gamification Transforming an annotation task into a (somewhat enjoyable) game can positively impact the annotation process (Chen et al., 2020; Goh and Lee, 2011; Mekler et al., 2013). First, enjoyable tasks make recruitment easier, as crowdworkers are more likely to accept them (Martin et al., 2014), and unpaid annotators, such as those in citizen science projects, may also prefer engaging and pleasant tasks. More importantly, gamification has shown promising results for linguistic annotation tasks in terms of annotation quantity and quality (Fort, 2016). When we collect annotations in a gamified setting, this approach can increase annotation output per person and overall data quality (Fort, 2016; Fort et al., 2018; Guillaume et al., 2016). In this context, competitive elements such as high scores or leaderboards may serve as motivating factors for players. However, setting up annotation games involves considerable financial and time costs (Fort, 2016). Therefore, gamification appears feasible and justifiable only in select contexts, such as repeated data collection phases.

Pre-Annotation Another design choice that shows promising results involves pre-annotating instances using either an automated or a human annotator. This approach can reduce annotation time and cost while maintaining quality. With this method, annotators do not view items in an unlabeled state but instead see a suggested label, which they must confirm or reject. However, while pre-annotations can accelerate the annotation process, they may also introduce bias by making annotators disproportionately likely to accept the pre-assigned label (Fort and Sagot,

2010; Fort and Claveau, 2012). These findings align with previous research that acknowledges the potential to improve data quality and reduce resource use, while also cautioning against the risk of bias (Dandapat et al., 2009; Mikulová et al., 2022; Rehbein et al., 2009). Other studies report benefits of pre-annotations in facilitating the annotation process without compromising data quality. Yet these studies often rely on inter-rater agreement as evaluation metric (Lingren et al., 2012, 2014). This reliance potentially biases quality metrics when different annotators encounter the same pre-annotated labels.

2.3.2. Data Composition

Thoughtful data composition decisions contribute to successful annotation projects and effective ML models. How we construct and organize annotated datasets shapes model performance, generalizability, and reliability. This subsection explores two key questions that researchers have addressed: how to effectively split annotated data into training and testing sets, and how many annotations each instance requires for optimal results.

Train-test Split The train–test split represents a foundational step in ML (Hastie et al., 2009), typically discussed in the context of ensuring fair and reliable model evaluation. However, we can also examine it from a data annotation perspective. In a study where annotators were asked to both generate new text examples and annotate them for training an NLP model, the researchers concluded that the same individuals should not be involved in both the creation of training and test data (Geva et al., 2019). In other words, the group responsible for generating and labeling training data should be distinct from the group annotating the test set, to avoid potential biases and overfitting. The main reason for this strict segregation of test and train data annotators is to prevent one (or very few) annotators from creating large shares of both train and test data. This overlap results in models that overfit (to that particular annotator's data) (Geva et al., 2019). Furthermore, another study evaluated models with extremely small datasets and costly annotations (here: autism classification and neuroimaging). The study observed decreased model accuracy that counterintuitively increased with larger sample sizes (Vabalas et al., 2019). Upon further evaluation, the authors found that they the training data was not split into test and train sets for models trained on extremely small datasets, a choice made to make the most use of every (sparse) annotation. However, this approach produced largely overfitted models that achieved high accuracy scores but did not generalize well outside the training data (Vabalas et al., 2019). This finding demonstrates that overlap between annotators in training and test sets can influence evaluation outcomes. While the study did not directly compare different train-test splitting strategies, it highlights potential biases when the same individuals contribute to both sets. High performance in controlled settings may not translate into real-world utility, which highlights the importance of careful evaluation practices.

Annotations per Instance In addition to concerns about the train-test split, designing an annotation task requires a decision about the number of annotations collected per instance (e.g., per image or phrase). Specifically, designers must estimate whether an additional annotation for an instance outweighs the benefit of annotating a new instance. This decision involves multiple parameters, including the availability of new instances, the costs per annotation of a (new) instance, task complexity, annotation quality, and the desired model outcome.

The trade-off between collecting an additional annotation versus a new instance depends on annotation quality and relative cost. When annotator quality is high, collecting a single label per instance is the most efficient strategy. In contrast, low annotator quality suggests collecting multiple annotations per instance (Sheng et al., 2008). However, when researchers use metrics like inter-annotator agreement to assess annotation quality, at least some instances need multiple annotators to label them.

Evidence suggests that when adding a new instance is cheap and annotations are costly, collecting an annotation for a new instance is more efficient than getting an additional annotation for an already labeled instance (Khetan et al., 2018). This becomes especially important when a model reaches a quality threshold. At this point, adding new annotated instances becomes most important for increasing the model's quality (Khetan et al., 2018). Other studies contradict this finding, showing that collecting many labels per instance performs better than collecting few labels across a larger set of instances (Gruber et al., 2024). This advantage stems mainly from leveraging the communicated uncertainty in multiple-labeled instances rather than hiding it in suboptimal aggregations such as majority votes. While this approach proves especially valuable for subjective annotation tasks, where multiple annotations help quantify annotator disagreement and uncertainty (Plank, 2022), uncertainty matters for objective tasks as well. Majority votes from multiple labels might offer greater robustness than single labels but wastefully discard the informational value of the uncertainy in the annotations (Gruber et al., 2023; Fleisig et al., 2023).

2.3.3. Monetary Incentives

In many cases, annotators receive payment for completing the task. This always applies to crowdworkers, but other groups such as student assistants or domain experts may also receive compensation. These groups sometimes earn payment specifically for each task or annotation session. The structure of monetary incentives likely influences annotation behavior and, ultimately, data and model quality. This subsection addresses annotator payment from two angles: the general wage level and the more complex design of flexible payment schemes.

Payment Level When designing annotation tasks, researchers must determine an appropriate wage level. Given a financial budget, higher wages result in fewer total annotations collected. However, insufficient wages also bring negative consequences that might offset the benefit of collecting more annotations. Tasks with inappropriate pay struggle to attract crowdsourced annotators, especially when competing with other tasks. Furthermore, even if annotators complete an underpaid task, Martin et al. (2014) observed that annotators generally consider exploiting (e.g., speeding through) poorly paid tasks more acceptable than doing so with properly paid ones. When crowdworkers have an approximate desired hourly wage in mind, they should be more likely to speed through underpaid tasks. Beyond influencing individual behavior, wage levels may also impact who chooses to participate. Without controlling for this selection effect, studies may conflate wage-related behavior with differences in the participant pool. Despite this theoretical reasoning, multiple studies conclude that higher wages do increase the quantity of work done (i.e., they facilitate recruitment) but not necessarily the quality of annotations (Auer et al., 2021; Buhrmester et al., 2011; Litman et al., 2015; Rogstadius et al., 2011; Vaughan, 2018; Wu et al., 2014; Ye et al., 2017).

Payment Flexibility How annotators receive payment influences their approach to annotation tasks. The first and most obvious decision involves choosing between a fixed payment per task or label and payment per time. In theory, neither option is strictly better. Fixed payments per task incentivize speeding and unthoughtful annotation, whereas payment per time incentivizes taking needless amounts of time per task without necessarily guaranteeing higher quality. Similar to many other paid tasks (such as responding to surveys), different strategies need assessment and validation. A more fine-grained approach to annotation incentives involves implementing performance-based bonus payments. This idea provides additional payments for high-quality annotation to improve annotation behavior. However, previous findings have been mixed. While one study observes improved data quality through performance-based payments (Ho et al., 2015), others cannot confirm this relationship (Lou et al., 2013; Shaw et al., 2011; Yin et al., 2013). This discrepancy may have multiple explanations, such as insufficient incentives relative to the required additional effort or simply that annotators already completed the task to their best ability. Fair compensation should ideally reflect annotators' effort or account for instance difficulty; otherwise, performance-based pay may unfairly penalize those assigned harder tasks.

While performance-based payments may sound promising for improving data quality, estimating annotator performance raises another fundamental problem. Without gold standard data available (which the annotation process often generates), we can only measure performance using imperfect indicators such as response time or agreement score with the majority label. If we knew parameters that perfectly measure annotator quality, the annotation would be obsolete. Ultimately, reducing annotators' leeway through incentives (or extreme guidelines) increases the degree to which the resulting dataset depends on the task requester. Even though task requesters often know the intended outcome, subjective annotation should aim to model disagreement between human annotators and uncertainty in labels.

2.3.4. Data Requirements

When designing annotation tasks, researchers must determine the required number of annotators and annotations through careful, data-driven planning. This process should ideally begin without considering constraints such as budget or annotator availability. These constraints can be incorporated later to define a realistic strategy. While sample size requirements remain flexible and may be adjusted, a priori estimations, such as power calculations, prove essential for a scientific and data-driven approach to model training. In contrast, collecting data only until performance plateaus or resources run out represents a suboptimal approach.

Required Sample Size Determining the optimal sample size for annotation projects requires balancing performance gains against collection costs. Data collection approaches that repeatedly predict the required sample size allow for flexible adjustments during annotation collection. We can achieve this by parallelizing data collection and model training processes and modeling the estimated performance curve (as measured by performance metrics, for example mean absolute error). Previous work has established theoretical foundations on how performance curves can help us estimate the value of individual data points in classifier models (Mukherjee et al., 2003). Based on the observed trajectory of the performance curve, we can predict the added value of an additional annotated data point and weigh it against the costs under the assumption of constant annotation quality. According to Figueroa et al. (2012), the performance curve generally follows

an "inverse power law" and modeling the learning curve is essential for finding the optimal sample size. They describe the common process to collecting annotated data as starting with "an initial number of samples in an ad hoc fashion to annotate data and train a model" (Figueroa et al., 2012, p. 9). In other words, rather than using predefined sample sizes, practitioners often begin with a small dataset and then gradually add more annotations if the model's performance falls short of the target. The authors argue that this strategy is "based on the vague but generally correct belief that performance will improve with a large sample size" (Figueroa et al., 2012, p. 9). Although an additional data point is unlikely to decrease model performance, we still need to weigh it against its costs. Therefore, the authors strongly advocate for modeling efforts and stress that the final strategy also depends on the required model performance and annotation costs. Active Learning could serve as an effective data collection framework to estimate the information gained by an annotation and, thereby, minimize the required sample size. This adaptive approach to sample size contrasts clearly with data collection processes in other applications, such as surveys or experiments, where practitioners mostly derive the sample size a priori (for example through power calculations). Additionally, increasing the sample size cannot resolve all data-related issues. If design-driven bias affects all instances, it will persist regardless of the number of observations (Gruber et al., 2023).

Required Positive Instances A slightly different approach to estimating the desired sample size involves focusing on the required positive instances, for example the number of positive instances of breast cancer on mammography results. Learning curves based on the number of positive instances in large datasets with very low positive rates (such as melanoma or other rare medical conditions) show that models can achieve high performance with relatively few positive examples (Richter and Khoshgoftaar, 2020). This study examined four datasets with over one million observations each and found that three required fewer than 2,500 positive instances to reach strong performance levels. These findings underline that the number of positive instances may serve as a better explanatory variable for model performance than the total sample size. The findings also suggest inspecting learning curves to make an informed judgment on sample requirements. Multiclass classification tasks add another layer of complexity to estimating sample size requirements from the number of instances per class.

Required Number of Annotators In addition to the previously discussed (demographic) distribution of annotators, an insufficient number of annotators may impact data and model quality. This issue affects annotation quality at two levels: individual instances benefit from multiple annotations, similar to seeking opinions from multiple doctors, and the overall dataset suffers when too few annotators dominate the collection process. Annotator constraints such as availability, costs, and quality should be weighed against the associated benefits to help task requesters estimate a target number of total annotators. In some domains, practitioners have already developed best practices, such as utilizing independent double coding followed by expert adjudication for occupation coding (Biemer and Caspar, 1994). The need to consider the number of annotators becomes clear from findings that, in many cases, a small number of annotators account for a disproportionately large share of the annotations (Geva et al., 2019). For example, in the Multi-Genre Natural Language Inference (MNLI) dataset, an eighth of the annotators produced around 90% of the total annotations. Since annotations nest within annotators (similar to survey interview responses nested within interviewers), allowing these large shares of annotations per individual

provides excessive leverage to single annotators and makes the training data prone to bias. Evidence supports this assumption: adding an annotator identifier as a model feature increased model performance across three of four examined datasets. Additionally, the clear individual component of annotations became evident when models trained to predict annotators based on their annotations performed well in the study. Furthermore, when annotators created new examples (to be annotated), a single-annotator trained model generalized worse to the test data of other annotators (Geva et al., 2019). These findings demonstrate that very small numbers of annotators or large shares of annotations per individual can bring unwanted consequences, especially in subjective tasks. Despite these risks, this overreliance is a frequently observed pattern (Kirk et al., 2023). While, especially with difficult or domain-specific tasks, the potential annotator pool is often small, analysts should at least evaluate the share of the total variability in the labels explained by the annotator IDs. Adding more annotators decreases each individual's impact on the model and may reduce the risk of a biased training dataset.

2.4. Automation

Since AI became widely adopted, particularly general-purpose language models, the field of data annotation has encountered a fundamental new question: to what extent can we automate annotation processes? Traditionally, annotation has relied heavily on human annotators, with automation limited to highly specific domains or requiring extensive pre-training. However, the general-purpose capabilities, low cost, and broad accessibility of modern language models have significantly transformed the annotation landscape.

Automated annotation is now feasible across nearly all data modalities, offering a fast and cost-effective alternative to manual efforts. Nonetheless, this shift also introduces new challenges. Among them are novel sources of bias that may be even less transparent and more difficult to interpret than human cognitive biases in annotation. Understanding and managing these risks remains critical, especially given the extremely rapid pace of technological advances.

This section explores recent developments in the field of annotation, focusing on both the opportunities and limitations of automated approaches. It also examines hybrid models, where human and AI annotators work collaboratively.

2.4.1. Automated Annotation

Opportunities Following the rapid rise in popularity of AI and language models, efforts quickly emerged to replace tasks traditionally carried out by humans with AI-based solutions, including in data annotation. Automated annotation has since demonstrated promising results across a range of domains and data modalities. These systems generate annotations that are often faster and cheaper, while maintaining comparable, or in some cases superior, quality to human annotations (Ding et al., 2023; Gilardi et al., 2023; Huang et al., 2023; Kuzman et al., 2023; Toney-Wails et al., 2024; Törnberg, 2023; Yu et al., 2024).

Within a relatively short period, general-purpose models have become a viable alternative to human annotators, particularly for tasks that demand lower levels of domain expertise and that practitioners often assign to less experienced annotators. In addition, as discussed in the section

on human annotation strategies, achieving human-level annotation quality may not always be necessary if the increased volume of annotations can compensate for the loss in individual annotation quality.

Challenges While automated annotation shows promising potential and often achieves results comparable to human annotation, it also introduces a wide range of challenges. Many of these remain poorly understood, and others may not yet be known or identified. Certain sources of bias and reductions in data quality may appear familiar and therefore easier to address, as they resemble known issues from human annotation processes. For instance, few-shot prompting uses examples for AI models, which resembles the example instances provided in human annotation guidelines. AI systems could also react to changes in task structure, such as the order in which instances are presented. Task complexity or burden, such as dealing with very long input documents, can influence data quality in ways that mirror human annotator fatigue. Even though these systems are not human, their exhibition of recognizable behavioral patterns makes certain sources of bias easier to detect. One example is positional bias – a systematic tendency to favor label options at certain positions, such as selecting the first option more frequently (Dominguez-Olmedo et al., 2024). Another example is the strong dependence on specific tasks and datasets when evaluating annotator quality, which complicates general assessments (Pangakis et al., 2023). In contrast, other challenges show less similarity to human annotation and therefore require new strategies. Issues such as the phrasing of prompts, the structure of the input context, the choice of sampling parameters, and the selection of specific model variants are all unique to automated systems (Pangakis and Wolken, 2025). Furthermore, it is not clear whether repeated annotations, possibly from different models, are necessary to ensure reliability (Egami et al., 2024). Similarly, determining whether certain instances in the annotation dataset were part of an LLM's original training data often proves difficult or impossible. Additional complications may arise from subtle and less visible factors, for example a model's inability to handle images with specific color distributions or resolutions. These issues affect most annotation tasks, but more fundamental concerns about automation emerge in the context of subjective annotation. Automated tasks that require personal judgment present unique challenges. For example, when assessing the trustworthiness of an individual or detecting hate speech in a piece of text, no clear answer exists regarding whose values and opinions the model reflects. Additionally, automated annotations can replicate problematic patterns and introduce systematic errors in subjective tasks. This behavior further reinforces concerns about their suitability and trustworthiness in contexts that rely on nuanced human judgment (Das et al., 2024; Felkner et al., 2024). Some emerging research has attempted to assign "personas" when prompting language models for use in annotation or prediction tasks, such as elections (Hu and Collier, 2024; von der Heyde et al., 2024). However, the foundational question of whether automated systems can legitimately represent subjective human judgment remains unresolved and requires resolution before widespread use of such systems in sensitive domains.

2.4.2. Human-AI Collaborative Annotation

Designing collaborative annotation workflows with human and automated annotators follows logically from the respective capabilities and limitations of both groups. Automated annotators can

generate annotations at low cost, with high speed, and at scale. Human annotators, in contrast, are well suited for tasks requiring domain-specific expertise and are represent subjective perspectives better than language models.

Opportunities Many studies propose hybrid annotation frameworks that emphasize the complementary strengths of humans and AI while mitigating their respective weaknesses (Gligorić et al., 2024; Li et al., 2023; Li, 2024; Wang et al., 2024). These studies often report promising outcomes. Hybrid annotation setups can accelerate the annotation process (Dreizin et al., 2023), reduce the workload for human annotators (van der Wal et al., 2021), enhance low-quality human annotations (Vădineanu et al., 2023), or lower the overall cost of annotation efforts (Beck et al., 2025b).

Challenges Despite these promising indications, interactive collaborative setups may introduce new threats to annotation quality. These risks add to the known challenges associated with both human and automated annotation. Because humans remain part of the loop in partially automated setups, we must again consider the full range of human biases. This becomes especially relevant when annotators interact with or respond to suggestions made by an AI system. Human annotators may over-rely on AI-generated annotations, a phenomenon called *automation bias* in the context of ADM. Alternatively they may display the opposite tendency, known as *algorithmic aversion*. Empirical evidence shows that automated pre-annotations can anchor human annotators, particularly under time constraints (Rastogi et al., 2022). Human-AI interactions can exacerbate annotation bias, with stronger effects than in human-human interactions (Glickman and Sharot, 2024). HCI and ADM research provide a growing body of transferable insights that explore methods for mitigating bias and improving collaborative dynamics (Dimara et al., 2019; Liu et al., 2024). However, within the domain of annotation research, investigations into such collaborative setups are still at an early stage. Additional forms of bias likely go unidentified and require systematic study.

2.5. Conclusion

2.5.1. Summary

In this background chapter, I outlined how various features and decisions within the annotation process can impact data and with that ML model quality. I structured this discussion around three dimensions: First, I discussed potentially biasing factors on the annotator side, such as demographic characteristics (for example, first language) and behavioral tendencies (such as misreporting). Second, I addressed strategic data collection decisions – from task design to data requirements – and their empirical evaluation. Finally, I examined the challenges and opportunities associated with the automation of annotation. This overview illustrates the complexity of working with annotated data and shows how both annotator-related and procedural factors can influence data quality, while pointing to possible ways to address them.

2.5.2. Contribution

While reviewing existing work and broader considerations, several gaps in the literature become evident and warrant attention. The articles in this dissertation each address open research questions, that aim to close specific gaps in the annotation literature.

Article 1 and Article 2 address a theoretical and methodological gap by applying concepts from survey methodology to the study of annotation. The articles answer the question of how sensitive human annotation is to small variations in task design and whether principles from survey methodology can be meaningfully applied to annotation tasks. This interdisciplinary transfer offers valuable insights into improving and understanding human annotation.

Article 3 addresses both a methodological and a substantive gap. The article explores how well different groups of annotators can perform an expert-level task. It also examines how quality relates to cost, and what pathways exist for partial automation. This study represents the first comprehensive comparison of multiple human and automated annotators for the same annotation task in remote sensing image classification. This domain presents unique annotation challenges due to the specialized domain knowledge required and the visual complexity of satellite imagery, which makes this comparative study particularly relevant.

Article 4 advances annotation practice by implementing a sequential pipeline that integrates automated and human annotators to reduce the demand on domain experts. It addresses the question of how LLMs can be effectively incorporated into a PDF-document annotation pipeline for an expert task, and what improvements are necessary to enhance such workflows. Moreover, the article provides a detailed account of the process, publishes the associated data, and aims to support the development of practically applicable annotation pipelines.

Article 5 unifies perspectives from HCI, ADM, and survey research to investigate bias in human-AI collaboration. The article explores how human annotators respond to AI-generated pre-annotations and how this interaction can introduce or amplify annotation bias. Through this approach, it addresses a timely issue in the context of annotation and automation, enabled by strong interdisciplinary theoretical foundations.

Part II.

Human Annotation Sensitivity

3. Improving Labeling Through Social Science Insights: Results and Research Agenda

Contributing article

Beck, J., Eckman, S., Chew, R., and Kreuter, F. (2022). Improving Labeling Through Social Science Insights: Results and Research Agenda. In *Proceedings of HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, pages 245–261.

Data repository

https://huggingface.co/datasets/soda-lmu/tweet-annotation-sensitivity-1

Author contributions

SE led the conceptualization in close collaboration with me. I designed and programmed the data collection tool and conducted the user study. SE ran the larger part of the analyses, I supported that part with complementary analyses. For the original draft, I wrote sections 1,2,3 and 5, SE wrote section 4. RC and FK contributed to the design of the experiments and reviewed the paper.

4. Order Effects in Annotation Tasks: Further Evidence of Annotation Sensitivity

Contributing article

Beck, J., Eckman, S., Ma, B., Chew, R., and Kreuter, F. (2024). Order Effects in Annotation Tasks: Further Evidence of Annotation Sensitivity. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 81–86.

Data repository

https://huggingface.co/datasets/soda-lmu/tweet-annotation-sensitivity-2

Author contributions

I developed the research question and conceptualized the project with SE consulted by BM, RC and FK. I curated the data and ran the analyses and visualization together with SE. I wrote the initial manuscript which was reviewed by all co-authors. BM designed the visual abstract.

Part III.

Automation in Annotation

5. Toward Integrating ChatGPT Into Satellite Image Annotation Workflows: A Comparison of Label Quality and Costs of Human and Automated Annotators

Contributing article

Beck, J., Kemeter, L. M., Dürrbeck, K., Abdalla, M. H. I., and Kreuter, F. (2025a). Toward Integrating ChatGPT Into Satellite Image Annotation Workflows: A Comparison of Label Quality and Costs of Human and Automated Annotators. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18, 4366–4381.

Author contributions

LMK and I jointly developed the research questions and drafted the manuscript. LMK and KD supervised and conducted the human annotation data collection and provided domain expertise in remote sensing. I conducted the automated data collection part and analyzed and visualized the data. MHIA conducted the cross-validation and reviewed the manuscript. KD and FK supervised the project and reviewed the manuscript. LMK and I contributed equally to this study.

6. Addressing Data Gaps in Sustainability Reporting: A Benchmark Dataset for Greenhouse Gas Emission Extraction

Contributing article

Beck, J., Steinberg, A., Dimmelmeier, A., Domenech Burin, L., Kormanyos, E., Fehr, M., and Schierholz, M. (2025b). Addressing data gaps in sustainability reporting: A benchmark dataset for greenhouse gas emission extraction. *Scientific Data*, 12, 1497.

Code repository

https://github.com/soda-lmu/gist-data-descriptor

Data repository

https://zenodo.org/records/15124118

Author contributions

Together with AS, I led the conceptualization of the study under supervision of MS (methodology) and AD (domain expertise). AS and I developed and administered the full data collection and annotation pipeline, curated, investigated and analyzed the data. We received support from LDB in the data collection and curation process. EK and MF contributed to the project as subject domain experts. MS supervised the project, I organized and coordinated the data collection phase. All authors contributed to the manuscript with larger contributions by AD, AS and me. AS and I contributed equally to this study. FK conceptualized the larger research project in which this paper is embedded and provided input to the study design.

Part IV.

Bias in Human-AI Collaborative Annotation

7. Bias in the Loop: How Humans Respond to AI-Generated Pre-Annotations

Contributing article

Beck, J., Eckman, S., and Kreuter, F. (2025c). Bias in the Loop: How Humans Respond to AI-Generated Pre-Annotations.

A revised version of this article is available on arXiv:2509.08514 and is currently under review at *Harvard Data Science Review*.

Author contributions

I received consulting by FK and SE in the derivation and formulation of research questions and hypotheses. I conceptualized, programmed, conducted and analyzed the user study including the visualization and writing of the manuscript. SE reviewed and edited the manuscript and provided substantial consulting with respect to data collection and evaluation methods.

Bias in the Loop: How Humans Respond to AI-Generated Pre-Annotations

JACOB BECK, LMU Munich, Department of Statistics & Munich Center for Machine Learning, Germany
STEPHANIE ECKMAN, University of Maryland, Social Data Science Center, USA
FRAUKE KREUTER, LMU Munich, Department of Statistics & Munich Center for Machine Learning, Germany University of Maryland, Social Data Science Center & Joint Program in Survey Methodology, USA

Annotation workflows are increasingly supported by artificial intelligence (AI)-generated pre-annotations to accelerate the process. However, these automated annotations can trigger cognitive biases that affect data quality. This study investigates how task design and annotator characteristics shape human responses to AI-generated pre-annotations. A Wizard of Oz-style experiment with crowdworkers performing emissions data annotation manipulated three factors: (1) pre-annotation quality (all correct or incorrect in the first three screens), (2) task burden (requiring or not requiring corrections), and (3) financial incentives via performance-based payments. Demographics, attitudes toward AI, and behavioral paradata were also collected. Performance was assessed using four metrics: accuracy, correction activity, overcorrection, and undercorrection. Results show that requiring corrections for incorrect pre-annotations reduced annotation activity and increased undercorrections. Financial incentives and early pre-annotation errors had no consistent effect on performance. Longer task engagement was linked to higher accuracy but also more overcorrections. Annotators more skeptical of AI were significantly more accurate, mainly due to lower undercorrection rates. These findings suggest that annotator characteristics, beyond standard demographics, significantly affect interactions with AI pre-annotations. Skepticism toward AI promotes more critical and accurate behavior. The results emphasize the importance of thoughtful task design: annotation pipelines combining AI and human input must consider how effort, incentives, and individual attitudes impact data quality. Even for objective annotation tasks, selecting diverse annotator samples and measuring relevant psychological traits can reduce bias and improve human-AI collaboration outcomes.

CCS Concepts: • **Human-centered computing → HCI design and evaluation methods**; **User studies**.

Additional Key Words and Phrases: Pre-annotation, Cognitive bias, Confirmation bias, Labeling accuracy

ACM Reference Format:
Jacob Beck, Stephanie Eckman, and Frauke Kreuter. 2025. Bias in the Loop: How Humans Respond to AI-Generated Pre-Annotations. 1, 1 (May 2025), 22 pages.

1 Introduction

The success of artificial intelligence (AI) models relies on high-quality annotated data for training, evaluation, and fine-tuning. Even large general-purpose models benefit from targeted fine-tuning to improve performance in specific domains or tasks. However, the landscape of data annotation has changed with the emergence of large language models (LLMs) capable of generating annotations at very low cost. Yet, LLM-generated annotations are imperfect. They are

Authors' Contact Information: Jacob Beck, jacob.beck@lmu.de, LMU Munich, Department of Statistics & and Munich Center for Machine Learning, Munich, Germany; Stephanie Eckman, steph@umd.edu, University of Maryland, Social Data Science Center, College Park, Maryland, USA; Frauke Kreuter, frauke.kreuter@lmu.de, LMU Munich, Department of Statistics & and Munich Center for Machine Learning, Munich, Germany and University of Maryland, Social Data Science Center & and Joint Program in Survey Methodology, College Park, Maryland, USA.

Manuscript submitted to ACM

prone to largely unknown and understudied inconsistencies and biases, making them unsuitable as standalone labels for high-stakes applications. This challenge raises the need for a resource-efficient and scalable approach that combines the advantages of human and automated annotators while maintaining high data quality.

One promising strategy is to rely on LLM-generated pre-annotations (PA), validating only a selected portion through a human annotator. This setup can take various forms, ranging from expert annotators to crowdworkers, and from randomly selecting annotation instances to prioritizing the most ambiguous or challenging ones [5, 28, 31]. Nevertheless, a common challenge across all approaches remains: human annotators are inherently prone to cognitive bias when interpreting information. Existing research around pre-annotations has largely neglected this danger of cognitive bias and focused on the time and cost saving aspect. However, other fields offer a rich body of literature on cognitive bias and bias mitigation strategies, which can be transferred to better understand biases in LLM-generated pre-annotations and how they might be addressed [12, 45, 57]. Pre-annotations can be approached through theoretical frameworks emerged from algorithmic decision-making (ADM), survey methodology or human-computer interaction (HCI) research. Importantly, Lyberg et al. [38] present a theoretical framework that outlines how human decision-making is shaped in a task called "dependent verification" (DV), which closely resembles the evaluation of pre-annotations. Within this framework, they propose four principles that explain psychosocial mechanisms in DV that can lead to systematic errors. These principles highlight a tendency not to correct all errors or to confirm existing information due to beliefs held about that information, phenomena we refer to collectively as **confirmation bias** [42, 45]. Building on this theory, we derive a set of hypotheses that guide our investigation into cognitive biases in human adjudication of pre-annotations.

In this study, we use a Wizard of Oz to test theoretical principles to understand crowdworkers' behavior when provided with pre-annotations. More precisely, we examine how confirmation bias arises and how manipulations of the incentive and workload structure, as well as the human annotator's beliefs about automation and AI affect these biases. Additionally, we closely investigate bias patterns conditioned on demographic information, annotation task device, and response time. We contribute to the discussion about LLM-generated annotations by identifying sources and drivers of bias and experimenting with practical solutions.

The remainder of the paper is structured as follows: First, we discuss related work, from which we derive and operationalize our hypotheses. Then, we illustrate our data, data collection and analytical methods followed by a presentation of the results. Ultimately, we discuss our findings, limitations and directions for future research.

2 Related Work

Human perception is influenced by a complex interplay of factors, including previously held beliefs and initial impressions. First impressions can significantly shape our perceptions and are often resistant to change, even when confronted with new information [22, 48, 63]. This phenomenon is closely linked to the well-established concept of confirmation bias, where individuals tend to favor information that aligns with their existing beliefs while disregarding or downplaying contradictory evidence [42, 45]. However, it is possible to mitigate this bias, and studies have explored strategies to avoid the confirmation trap by introducing specific stimuli or cognitive interventions [8, 50, 51, 62, 63].

2.1 Human Perception of Automated Decision-Making

Cognitive biases that shape human perception and decision-making also extend to automated decision-making (ADM). The way individuals judge algorithmic versus human decisions has been widely studied, yet the evidence remains mixed. A recurring theme is **algorithmic aversion**, where individuals tend to distrust AI-generated decisions and instead prefer decisions made by humans – either simply because they are made by humans [40], or due to a general skepticism

toward automation [60]. This aversion is influenced by factors such as knowledge about AI and task difficulty [27] and algorithmic transparency [29]. It can be reduced over time as people learn about an algorithm's performance [57]. Notably, people become more averse to algorithms after witnessing them make mistakes, even when the algorithm outperforms human decision-makers overall [11] and confidence is lost more quickly [29].

Other studies, however, find the opposite, that individuals favor automated decisions over human ones [9] and tend to over-rely on AI recommendations [21, 46]. They name this tendency **automation bias**. This effect is particularly strong when an algorithm has previously demonstrated superior performance [54]. When system strengths are encountered before weaknesses, automation bias skews user reliance and error rates [43]. The balance between algorithmic aversion and overreliance is shaped by multiple factors, including task difficulty [27], background knowledge of AI [21, 27], and demographics and psychological traits, including personality and familiarity with AI [39].

2.2 Pre-annotations - A special case of ADM

A significant shift in annotation practices is occurring with the rise of LLMs, which are increasingly used to generate automated annotations. For instance, GPT-based models can produce high-quality satellite image annotations [5]. However, automated annotations pose risks, particularly due to undetected bias, making sole reliance on such methods problematic. In contrast to human annotators, whose cognitive biases have been the subject of extensive research across disciplines, automated methods are still poorly understood in this regard. As a result, they are more likely to reproduce and reinforce existing, often undesirable, biases present in their training data [10, 16].

A promising, efficient, and scalable alternative to fully automated annotations is human validation of automated pre-annotations, where an AI-generated label is reviewed and, if necessary, corrected by a human. Automated pre-annotation can be considered a special case of an algorithmic decision that requires human judgment. However, this process differs from traditional ADM in several ways: (1) it is generally less cognitively demanding, as many pre-annotated instances are relatively straightforward and involve limited subjective judgment; (2) it is less consequential, since annotation decisions typically involve lower stakes than those in broader ADM contexts; and (3) the suggestions are clearly framed as pre-annotations, with human annotators explicitly aware that they can override the model's output.

Yet, despite these advantages, traditional sources of annotation bias remain relevant. Human judgment continues to be influenced by factors such as task structure (e.g., screen design and order effects), annotator demographics, and incentive structures [1, 3, 4, 15, 25, 30, 52, 53]. In addition to these established influences, hybrid annotation setups introduce new risks that arise specifically from the presence of automated suggestions. Annotators may undercorrect by overlooking mistakes, or overcorrect by unnecessarily changing correct suggestions. Research into the quality of data collected through human review of pre-annotations is scarce. While pre-annotations can reduce annotation time and costs [41], their effect on data quality and bias mitigation is underexplored. Some studies suggest that pre-annotations maintain high data quality, but these conclusions often rely on inter-rater agreement [35, 36], which may overestimate quality when multiple annotators are influenced by the same pre-annotations. Additionally, confirmation bias may lead annotators to accept suggestions without critical evaluation [17]. Together, these findings underscore a broader concern: while pre-annotations may improve efficiency, they can also introduce or reinforce biases.

2.3 Human-Computer Interaction

However, research relevant to this study extends beyond automated pre-annotations to a wide range of cooperative and interactive efforts between humans and automated agents. Numerous sources of bias have been empirically identified in HCI, many of which are not limited to interactions with large language models. This has become a prominent topic,

with significant momentum in recent years. On the one hand, promising results have been observed across various tasks, including text evaluation and annotation, as well as medical information extraction [33, 34, 58]. On the other hand, interacting with biased AI algorithms has been shown to amplify human biases, and research suggests that human-AI interactions can exacerbate biases more than human-human interactions [20]. Anchoring bias – the tendency to rely too heavily on an initial suggestion – increases in human-AI collaboration when decision-making time is limited [49]. In addition, humans often fail to recognize gender bias in robots trained on human-labeled data [24], highlighting the difficulty of detecting and correcting biases in AI-assisted systems. The scope of HCI research extends well beyond pre-annotation scenarios, offering valuable insights into how biases can emerge and influence human behavior in AI-assisted annotation workflows.

2.4 Theory from Dependent Verification

Since cognitive bias in pre-annotations remains largely unexplored in empirical research, the theoretical framework necessary to understand its impact can be borrowed from other disciplines. A framework developed for dependent verification in the context of coding survey responses [38] is especially relevant. Building on foundational work from social psychology, this framework outlines four principles that can lead to erroneous perception of a pre-annotator and, consequently, biased decision-making:

- **Principle 1: The human striving to reduce the cognitive working capacity** – People tend to minimize mental effort, which may lead them to accept pre-annotations without thorough evaluation.
- **Principle 2: Decisions based on heuristics** – Rather than carefully assessing each case, individuals rely on mental shortcuts, increasing the likelihood of systematic errors.
- **Principle 3: Psychosocial mechanisms such as liking or similarity can overrule cognitive reasoning** – Factors like familiarity, perceived competence, or implicit trust in a source can influence how pre-annotations are accepted or modified.
- **Principle 4: Humans expecting logic rather than randomness in the system** – Individuals assume that pre-annotations follow a structured pattern, leading to undue reliance on the system and result in undercorrection, even when errors are present.

These principles are highly transferable to pre-annotation tasks. However, they have yet to be empirically and experimentally tested in this specific context, highlighting a crucial research gap.

2.5 Possible Solutions

Approaches to mitigating bias often focus on automated solutions, such as in-process adjustments [19, 23] or post-processing and algorithmic refinements [18, 66]. However, there are also efforts on the human side, exploring ways to improve collaboration system design [12, 37].

In practical applications, particularly when working with crowdsourced labels, research has examined how incentive structures influence bias and motivation. Some studies suggest that performance-based payments (PBP) models can increase crowdworker engagement [26], while others find no such effect [55, 65]. Higher fixed payments do not consistently lead to improved label quality [2, 64], and task characteristics moderate these effects [61]. Overall, the evidence remains mixed, and PBP remains an underexplored strategy in bias mitigation.

3 Hypotheses

Following the discussion in the previous section, we draw on literature from social psychology, survey methodology, HCI, and ADM to formulate testable hypotheses on how confirmation bias influences human adjudication of automated pre-annotations. We translate the principles in Lyberg et al. [38] into hypotheses. Our performance metrics relate to accuracy, undercorrection, and overcorrection, examining key drivers and mediators of assigned corrections.

Principle 1: The human striving to reduce the cognitive working capacity

According to this principle, when the workload is equal across label options, no category should be favored over another. However, it is common for certain classes to come with an additional burden, such as providing a justification or answering a follow-up question. Human annotators are likely to recognize and adapt to this pattern, potentially favoring the class that reduces their workload. This kind of misreporting behavior is well known in survey research [13, 14, 32], and some evidence also exists for annotation tasks [7]. For these reasons, we formulate:

H1: An increase in the workload associated with correcting a pre-annotation leads to fewer corrections.

Principle 2: Decisions based on heuristics

The heuristic here is straightforward: the annotator's first impression is what matters. Since attention is generally assumed to be greatest at the beginning of a task [56], we hypothesize that the accuracy of the pre-annotations in the first three instances (out of ten) shapes participants' beliefs about the overall quality. These initial screens are intended to strongly influence any pre-existing heuristics or assumptions about the reliability of the pre-annotations. Repetition reinforces opinion formation, and a three-fold repetition has been shown to significantly impact this process [59]. Additionally, similar to a related Human-Robot Interaction context, we expect competence perception to develop fast and relatively resistant to change [47]. Therefore, we introduce strong tendencies in the first three screens and formulate:

H2: Displaying three incorrect vs. three correct pre-annotations in the first three instances affects the rates of undercorrections and overcorrections in subsequent instances.

Principle 3: Psychosocial mechanisms such as liking or similarity can overrule cognitive reasoning

According to Principle 3, human decisions and perceptions are shaped by their beliefs and attitudes towards the information and the source of the information. In the context of AI-generated content, individuals' attitudes about AI and automation are likely relevant. Therefore, we formulate:

H3: The human annotator‘s attitudes towards AI and automation affect accuracy and correction rates

Principle 4: Humans expect logic rather than randomness in the system

No testable hypothesis is derived from this principle, as it reflects a general cognitive tendency rather than a directly manipulable condition in our experimental design. While it may help explain undercorrections, such behavior alone does not confirm that this mechanism is at play. Moreover, within the scope of this study, we cannot directly observe whether annotators assume a logical pattern in the system, nor can we measure such expectations explicitly. Building on previous work in HCI, we operationalize and test an approach aimed at altering the monetary incentives of human annotators to mitigate the cognitive biases outlined in Principles 1-4:

H4: A performance-based bonus payment mitigates cognitive shortcutting and leads to higher annotation

accuracy.
All hypotheses and the study's approach were preregistered via the Open Science Framework (OSF).

4 Data and Methods

To test these four hypotheses, we conduct a Wizard of Oz user study using a factorial design, ensuring a high degree of control and enabling detailed measurement of the underlying patterns. The task involved extracting CO_2 emission values from tables found in company reports – an applied, real-world annotation scenario. A large sample of annotators completed this task under varying experimental conditions.

4.1 Data Collection

We collect annotations with crowdworking participants via Prolific in two steps. First, we fielded a survey about attitudes towards automation and AI. We use a six-item scale developed by [44] to measure the individual's attitudes towards AI and automation. Each question uses a seven point Likert-scale and we average responses over the six items. Survey respondents were required to be US residents and we requested a "representative" sample via Prolific, regarding age gender and ethnicity. The information published alongside the dataset used for this task, detailed in the next section and in [6], allowed us to estimate a design effect based on reported non-expert table annotations, which we used to conduct power calculations. To account for this design effect, the intended study required a sample size of 2,750 annotators. Anticipating some attrition, we admitted 3,200 crowdworkers to the survey, resulting in 3,187 complete cases.

Second, we created eight balanced strata of the survey respondents with respect to age, gender and ethnicity, variables provided by Prolific. These eight groups are then each invited to participate in one of the $2^3 = 8$ factorial experimental conditions. All 3,187 survey respondents were invited to the annotation task, from which up to 2,760 would have been admitted. We sent invitations to the annotation task one week after the close of the AI survey, to reduce the risk of contamination. This procedure of a "seemingly unrelated" survey is commonly used in survey practices. The median number of previously approved tasks on Prolific was 523 across our annotators, hence it is likely that after one week the automation is not salient anymore for most of the annotators.

Ultimately, 1,230 of the 3,187 invited survey respondents completed the annotation task, resulting in a response rate of 39%. To reach the required sample size and with that statistical power, we admitted additional annotators from Prolific who had not completed the survey (all of whom were US residents). The final sample contains 2,784 complete cases. Following the Prolific payment guidelines, the respondents were paid 0.75 GBP for the survey and 1.95 GBP for the annotation task.

4.2 Annotation Task

The annotation task involved extracting greenhouse gas (GHG) emissions data from tables in company reports. We choose this annotation task for multiple reasons:

(1) The availability of gold-standard labels, collected through a two-step expert annotation process [6], provides a reliable ground truth for evaluating annotation accuracy and bias.
(2) The task itself is well-defined: each instance has an objective true label, based on clear rules that do not rely on subjective judgment or domain-specific knowledge.
(3) The task is sufficiently complex and burdensome to trigger cognitive shortcutting behavior.

The human annotators were shown a table which contains GHG emission values on each annotation screen (Appendix 7). In addition, we present them with a pre-annotation for a given emission scope and reporting year (e.g., Scope 1 in 2020). The pre-annotations were framed as AI-generated, broadly referring to automated systems rather than specifically LLM-generated, and were in fact manually manipulated in a Wizard of Oz setup.

For more detail about the intial annotation procedure and rules as well as data and scripts, see [6]. Building on this data, we drew a stratified sample from a larger pool of gold-standard annotated tables, based on the agreement between non-expert annotators, whether either annotator was correct, and the type of label assigned to the instance. We manipulate the erroneous pre-annotations to contain four different types of errors to represent common errors observed in the LLM-generated annotations in [6]. Extracted emissions were incorrect due to one of the following: the wrong reporting year, the wrong scope, a spelling mistake/hallucination or wrong by the definition rules that were provided in the annotation tutorial and on the bottom of each screen. These errors were deliberately designed to mirror those encountered in [6], allowing us to investigate where human annotators are most likely to struggle. The Wizard of Oz design, in which the pre-annotations were manually crafted to simulate realistic automated annotations, enabled precise control over both the type and positioning of errors. Figure 1 illustrates the process of data collection. An example annotation screen is shown in Appendix 7.

After three tutorial instruction screens and two annotation examples, a negative and a positive one, all annotators saw the same 10 emissions tables. From these, three were randomly selected and always presented within the first three positions (in random order among themselves), while the remaining seven appeared afterward in a randomized sequence. Randomized order was ensured, but the annotation tool did not record the specific sequence in which each annotator viewed the screens. To test Hypothesis 2, we manipulated the pre-annotations of these first three tables as a ”treatment”: in one condition, all three were correctly pre-annotated; in the other, all three contained errors.
On each screen, the human annotators were shown an emission table along with a pre-annotation in the following format:

”The YEAR SCOPE emissions are VALUE, according to the AI. Is this correct?”

They were asked to judge whether the pre-annotation was correct or incorrect. As part of one experimental manipulation, annotators who selected “incorrect” were additionally required to provide a corrected value (see Appendix 7).

4.3 Experimental Conditions

The annotation task included three experimental manipulations, each with two levels, resulting in eight groups in a full factorial design. Annotators were randomly assigned to one of these condition combinations, as illustrated in Figure 1:

(1) **Asking for correct value if AI pre-annotation wrong**: If the human annotator classifies a pre-annotation as wrong they are asked vs. not asked for the correct value.
(2) **Error rate in first 3 instances**: The first three pre-annotations are all incorrect vs. all correct.
(3) **Performance-based payment**: For half of the annotators, a screen right before the start of the annotation tool, offered a bonus payment of 0.75*GBP* for the top 10% of annotators, judged by accuracy (defined below) [1].

[1] Immediately after the annotation data collection, we identified the most accurate annotators and approved the bonus payments.

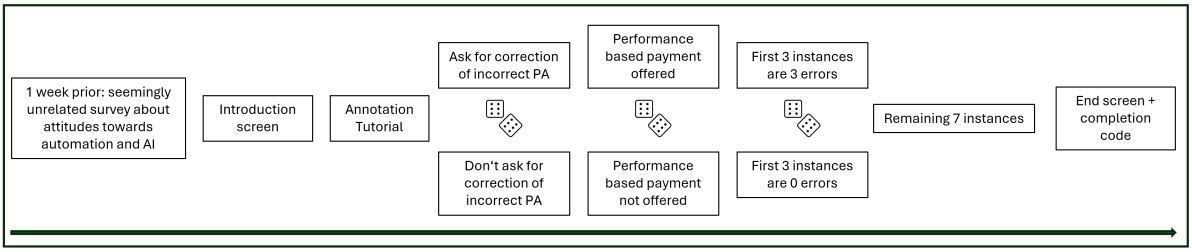

Fig. 1. Overview of the data collection process and experimental conditions.

Ultimately, a rich variety of data is available for each participant: the gold-standard annotations from [6], the previously collected attitudes on automation and AI, the demographic information from Prolific, the provided annotations, and task-related paradata such as time spent on each screen.

4.4 Evaluation Methods

Each human annotator saw N emission tables, each with a pre-annotation (PA). The annotator's task is to correct incorrect pre-annotations and leave correct pre-annotations unchanged. Thus, we define:

- N - Number of instances (emission tables) annotated.
- C - Number of correctly pre-annotated instances.
- I - Number of incorrectly pre-annotated instances.
- C_C - Number of correct PAs annotated as correct.
- C_O - Number of correct PAs falsely annotated as incorrect (overcorrection).
- I_C - Number of incorrect PAs annotated as incorrect.
- I_U - Number of incorrect PAs annotated as correct (undercorrection).

Every completed instance is annotated as correct or incorrect, thus $N = C + I$.

Performance Metrics

Accuracy. The accuracy measures the percentage of pre-annotations that were correctly handled:

$$\text{Accuracy} = \frac{C_C + I_C}{N} = \frac{C_C + I_C}{C + I}. \tag{1}$$

Overcorrection. Overcorrection occurs when the annotator indicates that a correct PA is not correct:

$$\text{Overcorrection} = \frac{C_O}{C}. \tag{2}$$

Undercorrection. Undercorrection occurs when the annotator indicates that an incorrect PA is correct:

$$\text{Undercorrection} = \frac{I_U}{I}. \tag{3}$$

We can represent the annotator's decisions in a 2×2 table as follows:

Annotator Decision	True Class: Correct PA	True Class: Incorrect PA
Annotate as correct	C_C (✓correct)	I_U (**undercorrection**)
Annotate as incorrect	C_O (**overcorrection**)	I_C (✓correct)

Table 1. Confusion Matrix for Annotator Performance

Since we deem the first three annotation screens, where the error rate is strongly manipulated to be either 0% or 100%, as experimental treatments, we consider just the remaining seven annotations for the calculation of the annotator performance metrics.

To test our four hypotheses, we regress our evaluation metrics on the experimental condition indicators, the individual annotator information such as demographic information or their stances towards automation, as well as annotation time. We run quasibinomial logistic regressions on the annotator level, that are well suited for modeling the accuracy metrics ranging between 0 and 1 as dependent variables. Unlike standard binomial models, quasibinomial models account for potential overdispersion, situations where the variability in the data exceeds what a standard binomial model would expect. This makes them particularly well suited in settings where additional variability is expected due to individual characteristics or experimental manipulations.

5 Results

This section is structured as follows: First, we address the hypotheses using descriptive analyses. We then examine patterns based on annotation time and explore differences between types of pre-annotation errors. Finally, we present regression models to gain a more detailed understanding of effects and interactions.

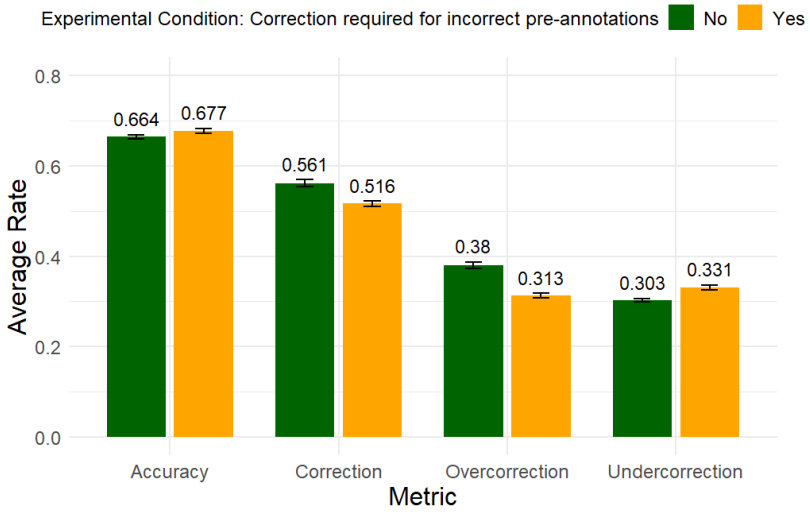

Fig. 2. Annotation performance metrics for annotators required vs. not required to correct wrong PAs

5.1 Descriptive Results

Hypothesis 1

As hypothesized, requiring corrections for PAs annotated as incorrect resulted in significantly fewer corrections, more undercorrections, and fewer overcorrections (Figure 2). Overall accuracy was significantly higher (68% vs 66%) when corrections were required ($p = 0.028$).

Hypothesis 2

The accuracy of the AI pre-annotation in the first three screens did not influence the human annotations of the subsequent seven pre-annotated emission tables (Figure 3). None of the performance metrics show meaningful differences in magnitude or statistical significance. While there is a slight increase in overcorrection after encountering three initial errors, this effect is minor and may be spurious. These findings provide strong evidence against H2, suggesting that human annotators do not form a strong and lasting impression of the pre-annotator at the start of the task.

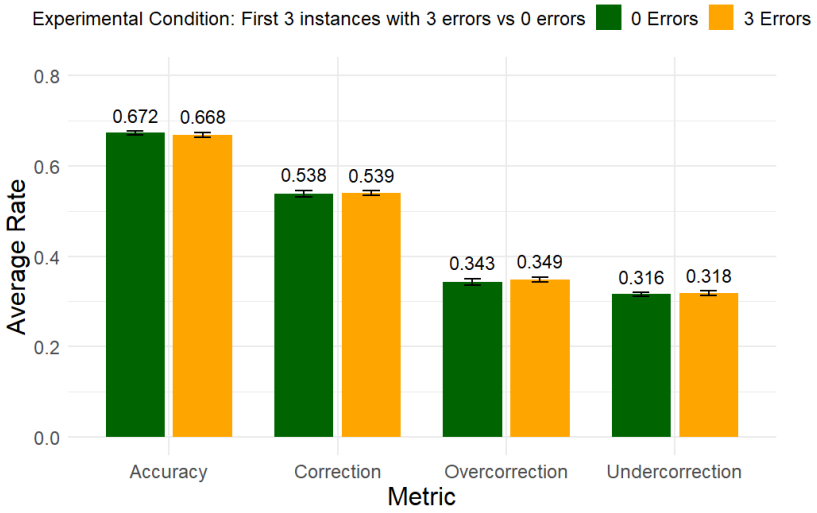

Fig. 3. Annotation performance metrics for annotators who saw 0 vs. 3 PA errors on the first three screens

Hypothesis 3

Figure 4 shows the four performance metrics subset by annotators' AI attitudes score, split into quartiles. Individuals with attitudes toward AI and automation in the lower two quartiles tend to correct AI-generated content more frequently than the upper quartiles, which also leads to higher rates of overcorrection. Conversely, those with a more favorable view of AI exhibit higher levels of undercorrection. Moreover, AI skeptics demonstrate greater overall accuracy, which appears to be driven by their lower tendency to undercorrect, indicating that they are less likely to trust AI-generated labels uncritically. The largest group, the annotators who did not participate in the initial survey, have similar accuracy, correction, and undercorrection rates to the survey participants in the higher AI liking quartiles. This group serves as a useful reference, as in typical annotation scenarios, where no prior survey is conducted, these are the annotators one would encounter, with their individual attitudes remaining unknown.

Even in this objective annotation task, we find evidence that psychosocial factors, such as attitudes towards AI, influence how annotators perceive and respond to pre-annotations. These individual-level characteristics are rarely measured and go beyond commonly reported demographics like age or gender. While it is generally assumed that such

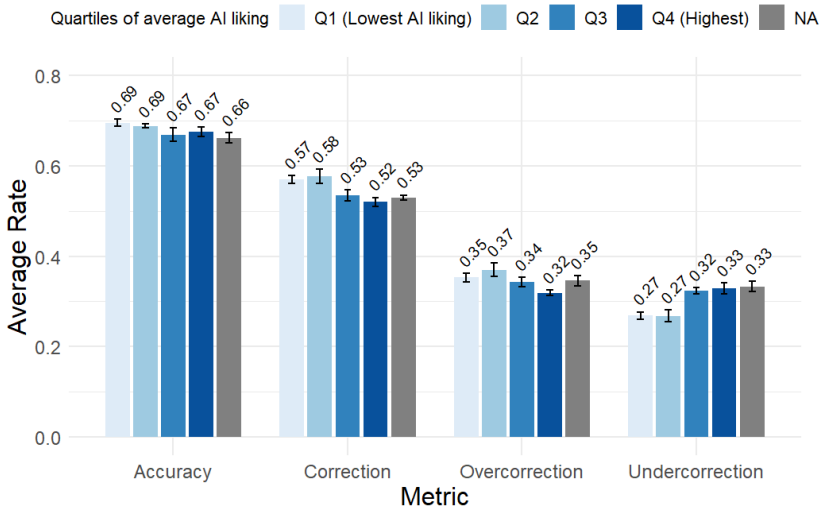

Fig. 4. Annotation performance metrics by Quartiles of AI Attitude Score

factors matter mainly for subjective tasks, our results show that even objective annotation outcomes depend on who performs the task, not just the task design itself.

Hypothesis 4

The promise of PBP does not meaningfully impact our performance metrics (Figure 5).

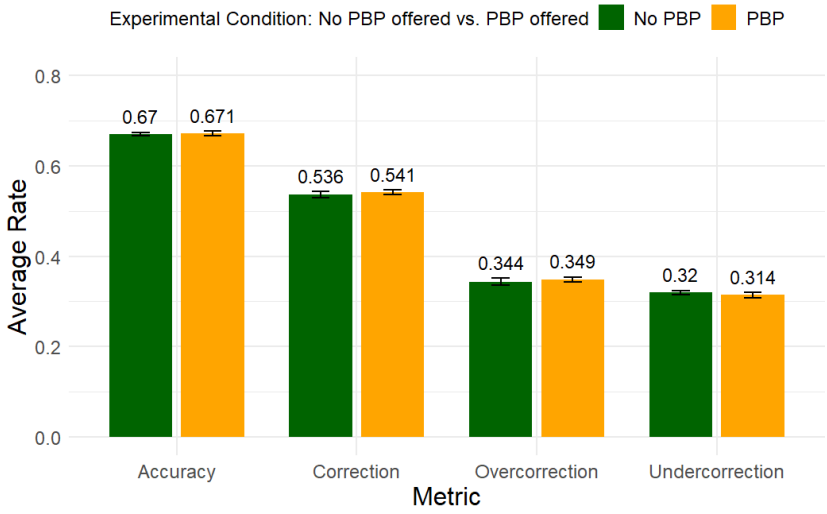

Fig. 5. Annotation performance metrics for annotators offered vs. not offered a PBP

While we observe a slight increase in corrections (and consequently overcorrections) when the PBP is available, the effect is too small to be meaningful. Notably, accuracy does not improve with the PBP, which could have several explanations. It is possible that performance is constrained not by motivation or incentives but by the inherent difficulty

of the task. Alternatively, the PBP used in this study (0.75*GBP*) may have been too low, or annotators may not have perceived it as realistically attainable (given that only the top 10% qualified).

While these interpretations remain speculative, the broader descriptive findings – such as the increased time spent when corrections were required and the rejection of H2 – suggest that human annotators were already putting forth their best effort, regardless of the PBP.

Response Time

Response time data can help shed light on how annotators engage with the task and whether timing patterns relate to annotation quality. However, neither the time spent on the whole task or just on the guidelines is correlated with accuracy ($r < 0.04$). While the average time spent on the first three screens is much higher (71 seconds) compared to the average for the remaining tables (48 seconds), on average annotators spent the same amount of time on a screen whether their assessment of the pre-annotation ended up being correct or incorrect (55 seconds). The same holds true for incorrect versus correct pre-annotations where we find no difference in average time spent on the screen (54 vs 55 seconds). Appendix 8 shows the average accuracy for all tables in relation to the average time spent on the screen. Two reports jump out with severely lower average accuracy. As the next paragraph will illustrate, these outliers are likely explained by the conceptual difficulty of correctly interpreting the table and evaluating the pre-annotation. When analyzing the overall trend without these two outliers, there is a slight negative relation between accuracy and time spent on the screen.

Error Types

The built-in errors in the Wizard of Oz pre-annotation screens varied in nature. Figure 6 illustrates the differences clearly.

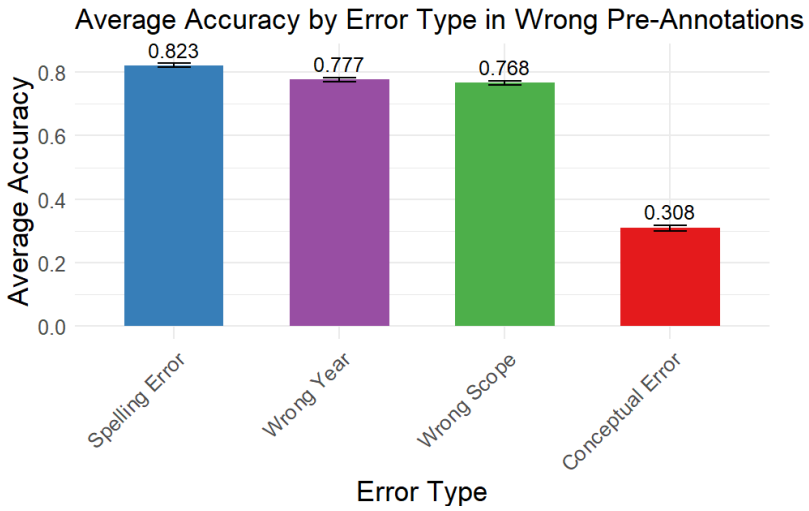

Fig. 6. Annotation accuracy by pre-annotation error type

Spelling mistakes, such as confusing two digits, were most frequently corrected, 82% of the time. Slightly more challenging, but still mostly corrected (around 77% of cases), were instances where the correct value was present in the table but located in a different cell, corresponding to a different year or scope. In contrast, we observe a striking drop in accuracy for screens where identifying the pre-annotation error required a conceptual understanding of the annotation

rules (31%). This pattern also holds for the Blackberry table (not included in the figure), where the pre-annotation was technically correct, but recognizing it as such required knowledge of market-based versus location-based Scope 2 emissions. Here, accuracy dropped to just 21%. These findings suggest that generalizable errors, like spelling mistakes or shifted cells, are easier to detect and may not require domain expertise. However, the low correction rates for conceptually challenging errors highlight the need for expert review in such cases, particularly when errors hinge on rule interpretation or domain-specific knowledge.

5.2 Modeling Analysis

To corroborate our descriptive findings and uncover potential interactions or hidden patterns not visible in the descriptive analysis, we estimated four quasibinomial logistic regression models. Table 2 presents the results, with one model for each outcome metric. The models include the randomly assigned experimental conditions along with all corresponding interaction effects. Additionally, we incorporate demographic variables sex, ethnicity, and age (grouped into three categories), as well as the custom scale measuring attitudes toward AI and automation.

The interpretation of the experimental conditions appears somewhat ambiguous. For the correction condition, the regression results confirm our descriptive findings: the condition that instructed participants to correct incorrect pre-annotations—thus requiring more effort—led to significantly fewer corrections. However, accuracy remained unchanged, as the reduced number of corrections was accompanied by a significantly lower rate of overcorrections. Neither the first three screen designs nor the PBP condition show significant regression coefficients for any of the four annotation performance metrics.

When examining the interaction effects of the experimental conditions, we find that participants who were asked to correct pre-annotations and were offered a PBP incentive corrected significantly fewer pre-annotations. This was accompanied by fewer overcorrections but more undercorrections. This outcome is somewhat difficult to interpret, as we initially hypothesized that the PBP incentive would lead to either higher accuracy or increased activity. In this case, however, the effect appears to be overshadowed by the additional effort required for performing corrections. In contrast, the interaction effect involving the condition that displayed three errors on the initial screens resulted in a significantly lower rate of undercorrection, seemingly driven by an overall increase in correction activity.

Additionally, we gain nuanced insights from variables we collected ourselves – specifically, annotators' attitudes toward AI and automation, as well as the time they spent on the task. The regression results reinforce our descriptive findings regarding annotators' self-reported attitudes toward AI and automation. A more favorable view of these abstract concepts was associated with lower accuracy and fewer corrections, likely driven by a significantly higher undercorrection rate. Notably, these patterns remain even after controlling for all other variables in the model.

Time spent on annotation (split into quartiles) was positively associated with accuracy, a higher number of corrections, and a lower rate of undercorrection. Higher total annotation time was also linked to more overcorrections. One possible explanation is that spending more time on a screen may lead annotators to perceive errors where none exist, or it may reflect uncertainty—causing them to err on the side of caution and flag more potential mistakes. However, these interpretations should be made with care.

Patterns based on demographic covariates are mixed. Male annotators showed a significantly higher accuracy rate and a lower overcorrection rate. Annotators identifying as Black had lower rates of accuracy and correction, and a higher undercorrection rate, when compared to the baseline group of Asian American annotators. A negative effect on the number of corrections, and thus an increase in undercorrection, was also observed for annotators identifying as "Other" or White, although these effects were not associated with significant changes in accuracy. In terms of

	Accuracy	Correction rate	OC rate	UC rate
Cond.: PBP	0.016	0.095	0.094	−0.119
	(0.052)	(0.061)	(0.068)	(0.097)
Cond.: First 3 errors	0.016	0.011	−0.008	−0.033
	(0.052)	(0.061)	(0.068)	(0.096)
Cond.: Correction	0.049	−0.146*	−0.239***	0.106
	(0.053)	(0.061)	(0.070)	(0.096)
Cond.: PBP × Cond.: First 3 errors	−0.061	−0.038	0.029	0.108
	(0.074)	(0.087)	(0.096)	(0.137)
Cond.: PBP × Cond.: Correction	−0.040	−0.213*	−0.226*	0.256+
	(0.074)	(0.086)	(0.098)	(0.135)
Cond.: First 3 errors × Cond.: Correction	−0.068	−0.062	−0.003	0.127
	(0.074)	(0.086)	(0.098)	(0.135)
Cond.: PBP × Cond.: First 3 errors × Cond.: Correction	0.136	0.202+	0.106	−0.337+
	(0.104)	(0.122)	(0.139)	(0.191)
AI Attitudes: Above or median (Ref.: Below median)	−0.121**	−0.200***	−0.117*	0.323***
	(0.041)	(0.047)	(0.053)	(0.075)
AI Attitudes: Missing	−0.120***	−0.124**	−0.020	0.246***
	(0.036)	(0.041)	(0.047)	(0.066)
Annotation Time: Q2 (Ref.: Q1)	0.273***	0.316***	0.086+	−0.551***
	(0.037)	(0.044)	(0.050)	(0.067)
Annotation Time: Q3	0.287***	0.397***	0.170***	−0.655***
	(0.037)	(0.044)	(0.050)	(0.069)
Annotation Time: Q4	0.281***	0.386***	0.165**	−0.634***
	(0.038)	(0.045)	(0.051)	(0.069)
Sex: Male (Ref.: Female)	0.063*	−0.008	−0.082*	−0.049
	(0.027)	(0.031)	(0.035)	(0.049)
Sex: Other	0.107	0.064	−0.043	−0.161
	(0.191)	(0.223)	(0.255)	(0.346)
Ethnicity: Black (Ref.: Asian)	−0.347***	−0.425***	−0.139+	0.743***
	(0.063)	(0.073)	(0.082)	(0.116)
Ethnicity: White	−0.099+	−0.143*	−0.067	0.247*
	(0.055)	(0.064)	(0.072)	(0.104)
Ethnicity: Mixed	−0.097	−0.123	−0.042	0.227+
	(0.072)	(0.084)	(0.094)	(0.135)
Ethnicity: Other	−0.113	−0.249**	−0.185+	0.371*
	(0.080)	(0.092)	(0.105)	(0.147)
Ethnicity: Missing	−0.136	0.020	0.173	0.115
	(0.152)	(0.180)	(0.196)	(0.289)
Age: 35-54 (Ref.: 18-34)	0.084**	0.136***	0.078+	−0.214***
	(0.030)	(0.035)	(0.040)	(0.055)
Age: 55+	0.058	0.079+	0.034	−0.132+
	(0.038)	(0.044)	(0.050)	(0.069)
Age: Missing	−0.592*	−1.374***	−1.115*	1.599**
	(0.294)	(0.373)	(0.441)	(0.551)
(Intercept)	0.638***	0.183*	−0.516***	−0.775***
	(0.073)	(0.085)	(0.095)	(0.135)
Num.Obs.	2738	2738	2738	2738
RMSE	0.15	0.19	0.20	0.26

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2. Quasi-binomial regression results

age, annotators aged 35–54 and 55+ corrected more screens and undercorrected less often compared to the baseline group (ages 18–34). The 35–54 age group also showed a significantly higher accuracy rate. We also observed large and significant coefficients for the "Missing" age group; however, these results cannot be meaningfully interpreted due to the small size of that subgroup.

6 Discussion

This study investigated how pre-annotation task designs and annotator characteristics influence cognitive biases and, ultimately, human annotation performance. We employed a factorial experimental design, manipulating the pre-annotator's error rate, task burden (by requiring corrections), and task reward (via PBP), while also collecting annotator-level characteristics and paradata. Notably, approximately half of the annotators had completed a survey assessing their attitudes toward automation and AI one week prior to the annotation task.

We evaluated annotation time and accuracy both descriptively and through regression models. Our findings indicate that annotation performance metrics were largely unaffected by the error rate in the first three screens or the presence of PBP incentives. However, when corrections were required, annotators were less likely to revise pre-annotations–suggesting that added effort reduced correction rates. This result replicates a robust finding in survey methods. Moreover, annotators who expressed greater skepticism toward automation and AI were more accurate in adjudicating pre-annotations, as they were less likely to overlook errors made by the automated system.

These findings offer insight into how task design and individual attitudes shape annotation behavior. The learnings should be carefully contextualized and discussed.

The absence of performance improvements under the PBP condition and the lack of association between annotator performance and the pre-annotation error rate in the initial screens suggest that crowdworkers were already exerting considerable effort. Moreover, the rejection of Hypothesis 2 suggests that annotators were not overly influenced by the first three screens and maintained consistent attention throughout the task. This interpretation is supported by several observations: some annotators sent direct messages expressing intrinsic motivation, and the time spent per screen was often higher than anticipated. Additionally, we observed table-specific differences in annotation accuracy. Such variation indicates a low prevalence of unwanted straightlining or speeding behaviors, as those would have produced uniform responses regardless of table content. For instance, the notably low accuracy on the Blackberry and JetBlue tables likely reflects the genuine difficulty or ambiguity of those specific items.

We also found that requiring annotators to provide a corrected value led to a measurable reduction in correction activity. If this mechanism is undesirable–for instance, if it discourages engagement with flawed pre-annotations–we recommend decoupling the task: one group of annotators could be assigned to judge the correctness of pre-annotations, while a separate group handles the correction of flagged cases. This would help ensure that annotator workload remains independent of the pre-annotation's assigned class, potentially mitigating effort-related biases. This result echos earlier findings that collecting more than one piece of information on one screen of the labeling instrument affects data quality [30].

Even in tasks that seem objective, like the one tested here, there are factors beyond standard demographics that influence how people annotate. While it is still important to choose suitable annotators (for example, only trained medical professionals should read X-ray images), other (normative) traits,such as attitudes, personal beliefs, or past experiences, can also matter. The ideal approach would be to measure these attitudes in advance and choose a diverse sample of annotators. However, this approach is not always feasible and we often do not know in advance what attitudes are relevant. A second-best approach is to routinely collect data on annotator demographics and aim for a large and

diverse group of annotators. Demographics may serve as proxy variables to capture differences in less visible traits that are linked to observable characteristics. For example, if individuals who are skeptical of AI are more likely to detect mistakes in AI-generated pre-annotations, this could become a problem. AI researchers, who are likely to hold more positive attitudes toward AI, may undercorrect errors if they conduct annotations themselves. Importantly, our findings suggest that disagreement between annotators and even low accuracy should not be dismissed as noise, but rather seen as a potentially valuable signal of instance difficulty or ambiguity.

Beyond annotation behavior, our results have some important implications for the domain of (semi-)automated extraction of GHG indicators, as examined in this study. Our findings show that even with pre-annotations and human adjudication, substantive errors can go unnoticed. As a result, this may unintentionally favor companies that report emissions in unclear, incomplete, or misleading ways.

These concerns, however, are not limited to AI skepticism or emissions data. Any domain or modality involving human-AI collaboration may be affected, each with its own challenges and relevant personal characteristics of annotators.

7 Limitations and Future Work

Some limitations of our study should be acknowledged. First, the task may not have been long enough to reveal the full effects over time. Differences caused by the experimental manipulations, especially those driven by fatigue or behavioral changes during a longer task, may not have become apparent. For example, the burden of providing a required correction might be perceived as more demanding as the task progresses, potentially increasing the likelihood of undercorrection. Second, correctly annotating the Blackberry report required a solid understanding of market-based and location-based Scope 2 emissions. Annotators appeared to struggle with assigning the correct labels, highlighting that not every annotator is suitable for every instance or even every task. As crowdworkers are unlikely to have the required domain knowledge, using a different report for this study could have been more informative. This observation connects to the issue of drawing generalizing conclusions from crowdworker studies. Patterns of motivation, bias, and performance may differ substantially for researchers, student assistants, volunteers, or domain experts.

Another limitation is that we could not analyze order effects, because the annotation tool did not track the random order in which tables were shown to each annotator. Previous work has shown that the order in which annotation tasks appear impacts the annotations given [4]. Additionally, we were unable to include a baseline condition without pre-annotations. It would have been valuable to understand how the annotation task would have played out without the influence of pre-annotated suggestions.

All of the discussed findings and illustrated limitations tie into the broader question of how to optimally set up a hybrid interactive pre-annotation pipeline. One possible setup could involve the use of LLMs in combination with expert evaluations. While such a system might avoid issues like the very low accuracy observed on two specific screens, it introduces new challenges. Experts are scarce, expensive, and their judgments often cover large portions of the data, which raises concerns about scalability and overreliance. In general, controlled experiments with annotator groups that are not crowdworkers could unravel how domain expertise, professional background, or institutional context relate to annotation behavior and bias.

Future research should focus on developing strategies for interactive annotation workflows that effectively balance the strengths and weaknesses of both human and automated annotators. This includes a deeper investigation into sources of bias, both in automated pre-annotations and in human cognitive processes. If incentives like PBP fail to improve outcomes, optimizing annotation quality will require attention to other key factors: the clarity and structure of guidelines, the quality and relevance of examples, task and screen design, the composition of the annotator sample,

and the interpretability of the annotation classes. As an extension of this study, purposefully manipulated annotation instances or attention checks could be placed throughout the task, not just at the beginning. This would allow researchers to assess whether annotators can detect them and potentially adapt the task's progression accordingly. Carefully designed experiments will help advance the field and support the development of evidence-based recommendations for human–AI collaborative annotation systems.

References

[1] 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* (Online, Nov. 2020), Association for Computational Linguistics, pp. 184–190.

[2] Auer, E. M., Behrend, T. S., Collmus, A. B., Landers, R. N., and Miles, A. F. Pay for performance, satisfaction and retention in longitudinal crowdsourced research. *Plos one 16*, 1 (2021), e0245460.

[3] Beck, J., Eckman, S., Chew, R., and Kreuter, F. Improving Labeling Through Social Science Insights: Results and Research Agenda. In *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence* (Cham, 2022), J. Y. C. Chen, G. Fragomeni, H. Degen, and S. Ntoa, Eds., Lecture Notes in Computer Science, Springer Nature Switzerland, pp. 245–261.

[4] Beck, J., Eckman, S., Ma, B., Chew, R., and Kreuter, F. Order effects in annotation tasks: Further evidence of annotation sensitivity. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)* (2024), pp. 81–86.

[5] Beck, J., Kemeter, L. M., Dürrbeck, K., Abdalla, M. H. I., and Kreuter, F. Toward integrating chatgpt into satellite image annotation workflows: A comparison of label quality and costs of human and automated annotators. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 18* (2025), 4366–4381.

[6] Beck, J., Steinberg, A., Dimmelmeier, A., Domenech Burin, L., Kormanyos, E., Fehr, M., and Schierholz, M. Addressing data gaps in sustainability reporting: A benchmark dataset for greenhouse gas emission extraction. *Under review* (2025). https://zenodo.org/records/15124118.

[7] Chandler, J. J., and Paolacci, G. Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science 8*, 5 (2017), 500–508.

[8] Chuanromanee, T., and Metoyer, R. A crowdsourced study of visual strategies for mitigating confirmation bias. In *2022 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (2022), IEEE, pp. 1–6.

[9] Chugunova, M., and Luhan, W. J. Ruled by robots: preference for algorithmic decision makers and perceptions of their choices. *Public Choice* (2024).

[10] Das, A., Zhang, Z., Hasan, N., Sarkar, S., Jamshidi, F., Bhattacharya, T., Rahgouy, M., Raychawdhary, N., Feng, D., Jain, V., et al. Investigating annotator bias in large language models for hate speech detection. In *Neurips Safe Generative AI Workshop 2024* (2024).

[11] Dietvorst, B. J., Simmons, J. P., and Massey, C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General 144*, 1 (2015), 114.

[12] Dimara, E., Bailly, G., Bezerianos, A., and Franconeri, S. Mitigating the Attraction Effect with Visualizations. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (Jan. 2019), 850–860.

[13] Duan, N., Alegria, M., Canino, G., McGuire, T. G., and Takeuchi, D. Survey Conditioning in Self-Reported Mental Health Service Use: Randomized Comparison of Alternative Instrument Formats. *Health Services Research*, 42 (2007), 890–907.

[14] Eckman, S., Kreuter, F., Kirchner, A., Jäckle, A., Presser, S., and Tourangeau, R. Assessing the Mechanisms of Misreporting to Filter Questions. *Public Opinion Quarterly 78*, 3 (2014), 721–733.

[15] Eckman, S., Plank, B., and Kreuter, F. Position: Insights from survey methodology can improve training data. In *International Conference on Machine Learning* (2024), PMLR, pp. 12268–12283.

[16] Felkner, V., Thompson, J., and May, J. Gpt is not an annotator: The necessity of human annotation in fairness benchmark construction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2024), pp. 14104–14115.

[17] Fort, K., and Claveau, V. Annotating Football Matches: Influence of the Source Medium on Manual Annotation. In *LREC - Eight International Conference on Language Resources and Evaluation* (Istanbul, Turkey, May 2012).

[18] Geyik, S. C., Ambler, S., and Kenthapadi, K. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining* (2019), pp. 2221–2231.

[19] Ghai, B., and Mueller, K. D-bias: A causality-based human-in-the-loop system for tackling algorithmic bias. *IEEE Transactions on Visualization and Computer Graphics 29*, 1 (2022), 473–482.

[20] Glickman, M., and Sharot, T. How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour 9*, 2 (Dec. 2024), 345–359.

[21] Goddard, K., Roudsari, A., and Wyatt, J. C. Automation bias: empirical results assessing influencing factors. *International journal of medical informatics 83*, 5 (2014), 368–375.

[22] Harris, C., Aarts, H., Fiedler, K., and Custers, R. Missing out by pursuing rewarding outcomes: Why initial biases can lead to persistent suboptimal choices. *Motivation Science 9*, 4 (2023), 288–297. Place: US Publisher: Educational Publishing Foundation.

[23] Heidrich, L., Slany, E., Scheele, S., and Schmid, U. Faircaipi: a combination of explanatory interactive and fair machine learning for human and machine bias reduction. *Machine learning and knowledge extraction 5*, 4 (2023), 1519–1538.

[24] Hitron, T., Megidish, B., Todress, E., Morag, N., and Erel, H. Ai bias in human-robot interaction: An evaluation of the risk in gender biased robots. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (2022), IEEE, pp. 1598–1605.

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

[26] Ho, C.-J., Slivkins, A., Suri, S., and Vaughan, J. W. Incentivizing High Quality Crowdwork. In *Proceedings of the 24th International Conference on World Wide Web* (Republic and Canton of Geneva, CHE, May 2015), WWW '15, International World Wide Web Conferences Steering Committee,

pp. 419–429.

[27] Horowitz, M. C., and Kahn, L. Bending the automation bias curve: a study of human and ai-based decision making in national security contexts. *International Studies Quarterly 68*, 2 (2024), sqae020.

[28] Huang, C., Deng, Y., Lei, W., Lv, J., and Dagan, I. Selective annotation via data allocation: These data should be triaged to experts for annotation rather than the model. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (2024), pp. 301–320.

[29] Jones-Jang, S. M., and Park, Y. J. How do people react to ai failure? automation bias, algorithmic aversion, and perceived controllability. *J. Comput. Mediat. Commun. 28* (2022).

[30] Kern, C., Eckman, S., Beck, J., Chew, R., Ma, B., and Kreuter, F. Annotation sensitivity: Training data collection methods affect model performance. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (2023), pp. 14874–14886.

[31] Krenzer, A., Makowski, K., Hekalo, A., Fitting, D., Troya, J., Zoller, W. G., Hann, A., and Puppe, F. Fast machine learning annotation in the medical domain: a semi-automated video annotation tool for gastroenterologists. *BioMedical Engineering OnLine 21*, 1 (2022), 33.

[32] 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 (2011), 88–104.

[33] Li, J. A comparative study on annotation quality of crowdsourcing and llm via label aggregation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2024), IEEE, pp. 6525–6529.

[34] Li, M., Shi, T., Ziems, C., Kan, M.-Y., Chen, N. F., Liu, Z., and Yang, D. CoAnnotating: Uncertainty-Guided Work Allocation between Human and Large Language Models for Data Annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (2023), pp. 1487–1505. arXiv:2310.15638 [cs].

[35] Lingren, T., Deléger, L., Molnár, K., Zhai, H., Meinzen-Derr, J., Kaiser, M., Stoutenborough, L., Li, Q., and Solti, I. Pre-annotating clinical notes and clinical trial announcements for gold standard corpus development: Evaluating the impact on annotation speed and potential bias. *2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology* (2012), 108–108.

[36] 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 : JAMIA 21*, 3 (May 2014), 406–413.

[37] Liu, Q., Jiang, H., Pan, Z., Han, Q., Peng, Z., and Li, Q. Biaseye: A bias-aware real-time interactive material screening system for impartial candidate assessment. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (2024), pp. 325–343.

[38] Lyberg, L., Eckman, S., Roos, M., and Kreuter, F. Cognitive aspects of dependent verification in survey operations. In *Diskussionsbeiträge und Materialien zur internationalen Berufsbildungszusammenarbeit* (Alexandria, VA, 2012), American Statistical Association, pp. 4310–4315.

[39] Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., and Smolander, K. What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change 175* (Feb. 2022), 121390.

[40] McGuire, J., and Cremer, D. D. Algorithms, leadership, and morality: why a mere human effect drives the preference for human over algorithmic leadership. *AI and Ethics 3* (2022), 601–618.

[41] Mikulová, M., Straka, M., Stepánek, J., těpánková, B., and Hajic, J. Quality and efficiency of manual annotation: Pre-annotation bias. *ArXiv abs/2306.09307* (2023).

[42] Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology 2*, 2 (1998), 175–220.

[43] Nourani, M., Roy, C., Block, J. E., Honeycutt, D. R., Rahman, T., Ragan, E., and Gogate, V. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces* (College Station TX USA, Apr. 2021), ACM, pp. 340–350.

[44] Novotny, M., Weber, W., Kern, C., and Kreuter, F. Measuring public opinion towards artificial intelligence: Development and validation of a general ai attitude short-scale. *Under Review* (2025).

[45] Oswald, M. E., and Grosjean, S. Confirmation bias. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory 79* (2004), 83.

[46] Packin, N. G. Consumer finance and ai: The death of second opinions? *Cyberspace Law eJournal* (2019).

[47] Paetzel, M., Perugia, G., and Castellano, G. The persistence of first impressions: The effect of repeated interactions on the perception of a social robot. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction* (2020), pp. 73–82.

[48] Rabin, M., and Schrag, J. L. First impressions matter: A model of confirmatory bias. *The quarterly journal of economics 114*, 1 (1999), 37–82.

[49] Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., and Tomsett, R. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction 6*, CSCW1 (2022), 1–22.

[50] Rieger, A., Draws, T., Theune, M., and Tintarev, N. This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias. In *Proceedings of the 32st ACM Conference on Hypertext and Social Media* (Virtual Event USA, Aug. 2021), ACM, pp. 189–199.

[51] Rieger, A., Draws, T., Theune, M., and Tintarev, N. Nudges to mitigate confirmation bias during web search on debated topics: Support vs. manipulation. *ACM Transactions on the Web 18*, 2 (2024), 1–27.

[52] Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., and Vukovic, M. An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. *Proceedings of the International AAAI Conference on Web and Social Media 5*, 1 (2011), 321–328. Number: 1.

[53] 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 *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2022), pp. 5884–5906.

[54] Saragih, M. S. B., and Morrison, B. W. The effect of past algorithmic performance and decision significance on algorithmic advice acceptance. *International Journal of Human–Computer Interaction 38* (2021), 1228 – 1237.

[55] Shaw, A. D., Horton, J. J., and Chen, D. L. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (New York, NY, USA, Mar. 2011), CSCW '11, Association for Computing Machinery, pp. 275–284.

[56] Tu, J., Yu, G., Wang, J., Domeniconi, C., and Zhang, X. Attention-aware answers of the crowd. In *Proceedings of the 2020 SIAM International Conference on Data Mining* (2020), SIAM, pp. 451–459.

[57] Turel, O., and Kalhan, S. Prejudiced against the machine? implicit associations and the transience of algorithm aversion. *MIS Q. 47* (2023), 1369–1394.

[58] Wang, X., Kim, H., Rahman, S., Mitra, K., and Miao, Z. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–21.

[59] Weiss, R. F. Repetition of Persuasion. *Psychological Reports 25*, 2 (Oct. 1969), 669–670. Publisher: SAGE Publications Inc.

[60] Wesche, J. S., Hennig, F., Kollhed, C. S., Quade, J., Kluge, S., and Sonderegger, A. People's reactions to decisions by human vs. algorithmic decision-makers: the role of explanations and type of selection tests. *European Journal of Work and Organizational Psychology 33* (2022), 146 – 157.

[61] Wu, H., Corney, J., and Grant, M. Relationship between quality and payment in crowdsourced design. In *Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (2014), IEEE, pp. 499–504.

[62] Wyer, N. A. You Never Get a Second Chance to Make a First (Implicit) Impression: The Role of Elaboration in the Formation and Revision of Implicit Impressions. *Social Cognition 28*, 1 (Feb. 2010), 1–19.

[63] Ybarra, O. When first impressions don't last: The role of isolation and adaptation processes in the revision of evaluative impressions. *Social Cognition 19*, 5 (2001), 491–520.

[64] Ye, T., You, S., and Robert Jr, L. When does more money work? examining the role of perceived fairness in pay on the performance quality of crowdworkers. In *Proceedings of the International AAAI Conference on Web and Social Media* (2017), vol. 11, pp. 327–336.

[65] Yin, M., Chen, Y., and Sun, Y.-A. The Effects of Performance-Contingent Financial Incentives in Online Labor Markets. *Proceedings of the AAAI Conference on Artificial Intelligence 27*, 1 (June 2013), 1191–1197. Number: 1.

[66] Zhu, Z., Wang, J., and Caverlee, J. Measuring and mitigating item under-recommendation bias in personalized ranking systems. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (2020), pp. 449–458.

8 Appendix

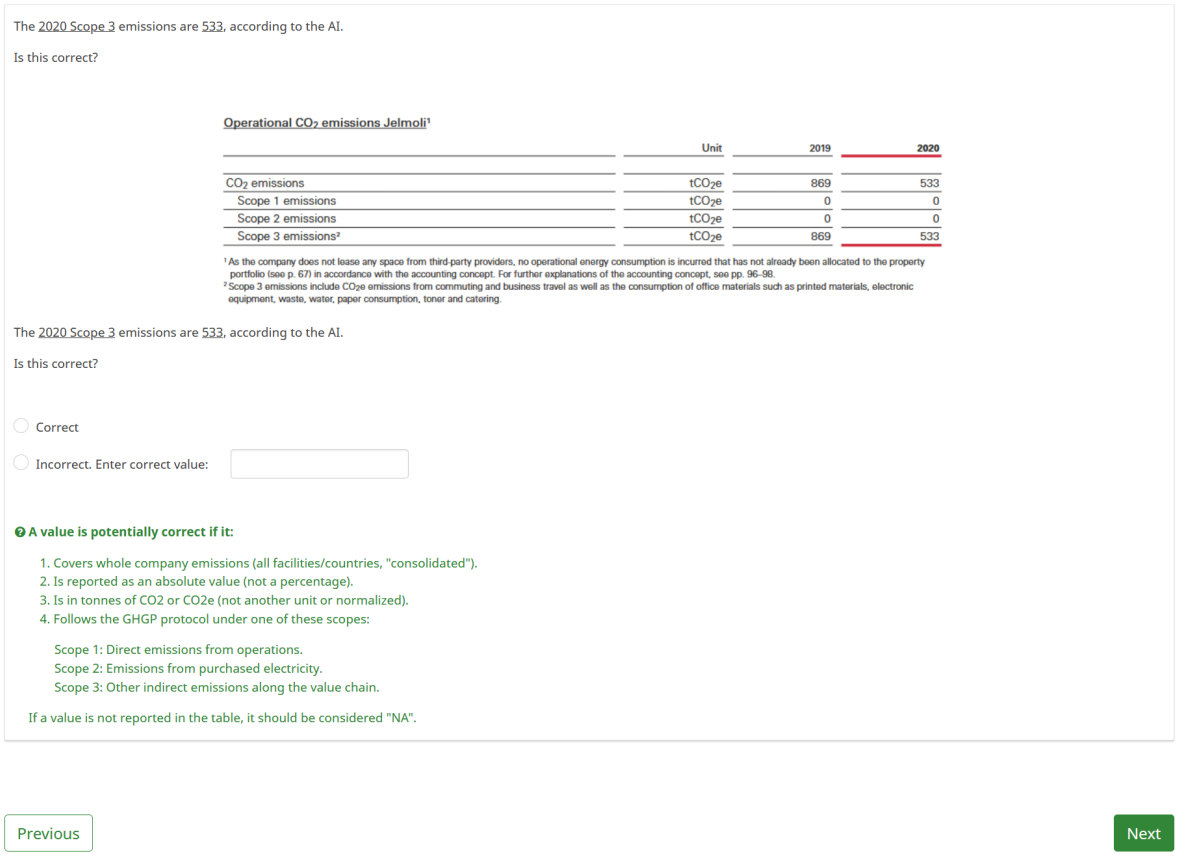

Fig. 7. Example screenshot of annotation instrument

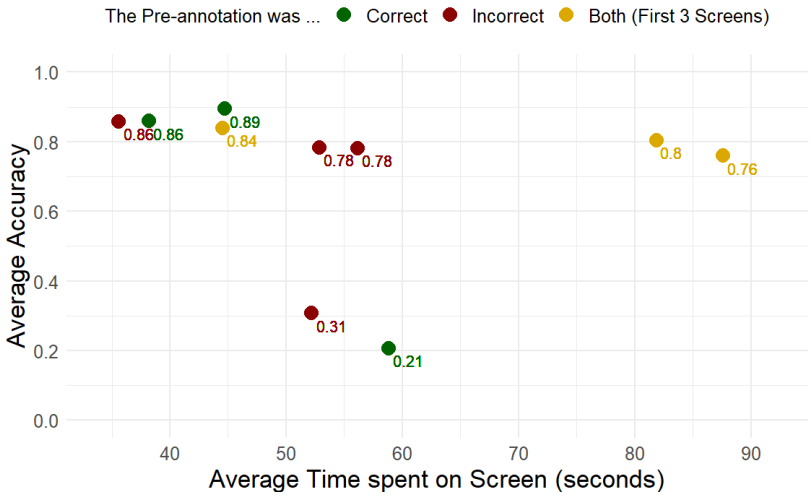

Fig. 8. Average accuracy and time spent on each screen

Part V.

Concluding Remarks

Summary The quality of annotated data sets the upper bound for model performance; better labels enable better models. In this dissertation, I took a "science of data" perspective to examine how to improve data annotation and what factors impact its quality. Part II. demonstrated that human annotation is highly sensitive to minor manipulations in task design and highlighted the wide range of design choices involved, whether in task setup or in annotator selection and deployment. Part III. showed how automation elements improve annotation by reducing the need for expert involvement and by saving time and cost. I assessed how well partial automation fits into annotation pipelines and how effectively it performs. Part IV. investigated whether collaborative annotation workflows between humans and automated systems introduce new forms of bias. These findings carry important implications for the field and for anyone conducting annotation work.

Acknowledging Sensitivity Understanding that annotators do not assign labels independently of the data collection context has important implications for how annotation tasks should be designed. Whether errors occur randomly or systematically, for example due to annotator characteristics, several strategies can mitigate their impact. Many of these approaches are often feasible to implement. Most importantly, before setting up data collection, one should take the time to reflect on the task and its circumstances: What is the intended outcome of the annotation and the downstream application? What is the degree of subjectivity involved? What are the options regarding potential annotators? Can automation assist the process? Matching task requirements to available resources is a crucial first step, yet one that people often undertake without sufficient reflection. For instance, the degree of task subjectivity can inform whether one should minimize annotator heterogeneity, through strict guidelines, or embrace it by collecting multiple labels per instance. Moreover, annotation setups often allow for measuring potentially informative indicators, such as annotator demographics, annotation time and order, or measures of uncertainty and disagreement. Including such metrics can help identify unwanted imbalances, support post-processing efforts, and strengthen claims regarding dataset quality and its suitability for reuse. Other potentially helpful solutions, such as experimenting with alternative task versions, involving a large group of diverse annotators, or conducting a preliminary survey, prove often infeasible in practice. This is understandable; however, when similar annotation tasks occur repeatedly, it may be worthwhile to experiment with varied task designs and evaluate the resulting annotation quality.

Automating Annotation In line with other research, I demonstrated that integrating elements of automation into annotation workflows shows promise. I presented two examples of how automated annotations can integrate into hybrid setups. Given the current landscape of annotation methods and automation capabilities, hybrid approaches appear particularly well-suited, and their popularity will likely increase. Current general-purpose models can already match the performance of lower-quality annotation approaches in some cases, such as large-scale crowdsourcing. Automated labels often generate more quickly, at lower cost[1], and with greater consistency and control. However, in domains that require extensive subject-matter expertise, completely removing human experts from the annotation pipeline requires great caution – if we consider it at all. At minimum, human annotators should validate a subset of the automated annotations. This caution applies even more strongly to subjective annotation tasks, where automated systems may not adequately

[1] Notably, the apparent cost-efficiency of LLMs does not account for their environmental impact. Training and deploying such models require substantial computational resources, and their true cost is currently not reflected in market prices.

reflect human ambiguity, values, or opinions. We should be especially mindful of the influence we grant to language models in shaping or reflecting societal norms and perspectives. Overreliance on such systems poses risks, as they tend to reproduce unwanted patterns and may cause significant harm when people trust them uncritically.

Human-AI Collaborative Annotation Combining human and automated annotators may serve as a transitional solution, but it also holds potential as a lasting practice in annotation workflows. However, we must remain aware that hybrid setups introduce new forms of bias, particularly through human-AI interaction. The results indicated that even in an objective annotation task, individuals' general attitudes toward automation influenced human behavior, when reviewing AI-generated pre-annotations. This observation highlights the need for further investigation to better understand and improve these promising collaborative setups.

Flexible designs offer substantial potential to reduce the resource intensity of annotation data collection. For example, an annotation task could always collect two independent labels per instance, adding a third only if the initial annotators disagree. These flexible approaches are well-suited to incorporate automated annotations, for example by using automation for easy-to-label instances or to generate pre-annotations. Adaptive strategies help balance quality assurance with efficiency, making annotation pipelines more scalable and cost-effective while maintaining high quality.

Finding Balance The field of data annotation must find a balanced path between overemphasizing the omnipresence of errors and assuming that models can correct for all imperfections in the data. While it is unrealistic to eliminate all biases during the training process, the mere presence of bias in annotated data does not render it unusable. Simply calling out bias in every instance does little to advance the field in practice, just as ignoring annotation bias is likely to cause models to hit a performance ceiling in the long run. The field needs a synthesis of perspectives: the ML engineering approach of "making it work" must be complemented by a "science of data" perspective – one that places data at the center of analysis, rather than treating it as a raw material to be simply processed.

Contributing Articles

Beck, J., Eckman, S., Chew, R., and Kreuter, F. (2022). Improving Labeling Through Social Science Insights: Results and Research Agenda. In *Proceedings of HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, pages 245–261.

Beck, J. (2023). Quality Aspects of Annotated Data – A Research Synthesis. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 17(3), 331–353.

Beck, J., Eckman, S., Ma, B., Chew, R., and Kreuter, F. (2024). Order Effects in Annotation Tasks: Further Evidence of Annotation Sensitivity. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 81–86.

Beck, J., Kemeter, L. M., Dürrbeck, K., Abdalla, M. H. I., and Kreuter, F. (2025a). Toward Integrating ChatGPT Into Satellite Image Annotation Workflows: A Comparison of Label Quality and Costs of Human and Automated Annotators. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18, 4366–4381.

Beck, J., Steinberg, A., Dimmelmeier, A., Domenech Burin, L., Kormanyos, E., Fehr, M., and Schierholz, M. (2025b). Addressing data gaps in sustainability reporting: A benchmark dataset for greenhouse gas emission extraction. *Scientific Data*, 12, 1497.

Beck, J., Eckman, S., and Kreuter, F. (2025c). Bias in the Loop: How Humans Respond to AI-Generated Pre-Annotations.

References

Aguda, T. D., Siddagangappa, S., Kochkina, E., Kaur, S., Wang, D., and Smiley, C. (2024). Large language models as financial data annotators: A study on effectiveness and efficiency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10124–10145, Torino, Italia. ELRA and ICCL.

Al Kuwatly, H., Wich, M., and Groh, G. (2020). Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Ang, C. S., Bobrowicz, A., Schiano, D. J., and Nardi, B. (2013). Data in the wild: Some reflections. *Interactions*, 20(2): 39–43.

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*, CHI '12, pages 2925–2934, New York, NY, USA. Association for Computing Machinery.

Arhin, K., Baldini, I., Wei, D., Ramamurthy, K. N., and Singh, M. (2021). Ground-truth, whose truth? – Examining the challenges with annotating toxic text datasets. arXiv preprint arXiv:2112.03529.

Aroyo, L. and Welty, C. (2015). Truth is a Lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1): 15–24.

Auer, E. M., Behrend, T. S., Collmus, A. B., Landers, R. N., and Miles, A. F. (2021). Pay for performance, satisfaction and retention in longitudinal crowdsourced research. *PLOS One*, 16(1): e0245460.

Beatty, P. C. and Willis, G. B. (2007). Research synthesis: The practice of Cognitive Interviewing. *Public Opinion Quarterly*, 71(2): 287–311.

Beck, J., Eckman, S., and Kreuter, F. (2025a). Bias in the loop: How humans respond to AI-generated pre-annotations. *Manuscript submitted for publication.*

Beck, J. (2023). Quality aspects of annotated data: A research synthesis. *AStA Wirtschafts-und Sozialstatistisches Archiv*, 17(3): 331–353.

Beck, J., Eckman, S., Chew, R., and Kreuter, F. (2022). Improving labeling through social science insights: Results and research agenda. In *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, Lecture Notes in Computer Science, pages 245–261, Cham. Springer Nature Switzerland.

Beck, J., Eckman, S., Ma, B., Chew, R., and Kreuter, F. (2024). Order effects in annotation tasks: Further evidence of annotation sensitivity. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 81–86.

Beck, J., Kemeter, L. M., Dürrbeck, K., Abdalla, M. H. I., and Kreuter, F. (2025b). Toward integrating ChatGPT into satellite image annotation workflows. A comparison of label quality and costs of human and automated annotators. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Belletti, C., Erdsiek, D., Laitenberger, U., and Tubaro, P. (2021). Crowdworking in France and Germany. Technical report, ZEW Expert Brief.

Berinsky, A. J., Huber, G. A., and Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(3): 351–368.

Biemer, P. and Caspar, R. (1994). Continuous quality improvement for survey operations: Some general principles and applications. *Journal of Official Statistics*, 10(3): 307–326.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74(5): 817–848.

Biemer, P. P., de Leeuw, E. D., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., Tucker, N. C., and West, B. T. (2017). *Total survey error in practice*. John Wiley & Sons.

Biester, L., Sharma, V., Kazemi, A., Deng, N., Wilson, S., and Mihalcea, R. (2022). Analyzing the effects of annotator gender across NLP tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 10–19, Marseille, France. European Language Resources Association.

Binns, R., Veale, M., Van Kleek, M., and Shadbolt, N. (2017). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II 9*, pages 405–415. Springer.

Bless, H. and Schwarz, N. (2010). Chapter 6 - Mental construal and the emergence of Assimilation and Contrast effects: The Inclusion/Exclusion model. In *Advances in Experimental Social Psychology*, volume 42 of *Advances in Experimental Social Psychology*, pages 319–373. Academic Press.

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, Data? *Perspectives on Psychological Science*, 6(1): 3–5.

Bui, M. D., Wense, K. V. D., and Lauscher, A. (2025). Multi^{3}Hate: Multimodal, multilingual, and multicultural hate speech detection with vision–language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9714–9731, Albuquerque, New Mexico. Association for Computational Linguistics.

Cefkin, M., Anya, O., Dill, S., Moore, R., Stucky, S., and Omokaro, O. (2014). Back to the future of organizational work: Crowdsourcing and digital work marketplaces. In *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work &*

social computing, CSCW Companion '14, pages 313–316, New York, NY, USA. Association for Computing Machinery.

Chandler, D. and Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, 90: 123–133.

Chandler, J. J. and Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, 8(5): 500–508.

Chen, C.-M., Li, M.-C., and Chen, T.-C. (2020). A web-based collaborative reading annotation system with gamification mechanisms to improve reading performance. *Computers & Education*, 144: 103697.

Chen, Y. and Joo, J. (2021). Understanding and mitigating annotation bias in facial expression recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14960–14971, Montreal, QC, Canada. IEEE.

Cowan, G. and Khatchadourian, D. (2003). Empathy, ways of knowing, and interdependence as mediators of gender differences in attitudes toward hate speech and freedom of speech. *Psychology of Women Quarterly*, 27(4): 300–308.

Dandapat, S., Biswas, P., Choudhury, M., and Bali, K. (2009). Complex linguistic annotation – No easy way out! A case from Bangla and Hindi POS labeling tasks. In *Proceedings of the Third Linguistic Annotation Workshop*, ACL-IJCNLP '09, pages 10–18, USA. Association for Computational Linguistics.

Das, A., Zhang, Z., Hasan, N., Sarkar, S., Jamshidi, F., Bhattacharya, T., Rahgouy, M., Raychawdhary, N., Feng, D., Jain, V., et al. (2024). Investigating annotator bias in large language models for hate speech detection. In *Neurips Safe Generative AI Workshop 2024*.

Davani, A. M., Atari, M., Kennedy, B., and Dehghani, M. (2023). Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11: 300–319.

Davidson, T. and Bhattacharya, D. (2020). Examining racial bias in an online abuse corpus with structural topic modeling. arXiv preprint arXiv:2005.13041.

Dimara, E., Bailly, G., Bezerianos, A., and Franconeri, S. (2019). Mitigating the Attraction effect with visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1): 850–860.

Ding, B., Qin, C., Liu, L., Chia, Y. K., Li, B., Joty, S., and Bing, L. (2023). Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195.

Ding, Y., You, J., Machulla, T.-K., Jacobs, J., Sen, P., and Höllerer, T. (2022). Impact of annotator demographics on sentiment dataset labeling. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–22.

Dominguez-Olmedo, R., Hardt, M., and Mendler-Dünner, C. (2024). Questioning the survey responses of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 45850–45878.

Dreizin, D., Zhang, L., Sarkar, N., Bodanapally, U. K., Li, G., Hu, J., Chen, H., Khedr, M., Khetan, U., Campbell, P., et al. (2023). Accelerating voxelwise annotation of cross-sectional imaging through AI collaborative labeling with quality assurance and bias mitigation. *Frontiers in radiology*, 3: 1202412.

Dumitrache, A., Aroyo, L., and Welty, C. (2015). Achieving expert-level annotation quality with crowdtruth. In *Proceedings of BDM2I Workshop, ISWC*.

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

Egami, N., Hinck, M., Stewart, B. M., and Wei, H. (2024). Using large language model annotations for the social sciences: A general framework of using predicted variables in downstream analyses. Manuscript available at `https://naokiegami.com/paper/dsl_ss.pdf`.

Eickhoff, C. (2018). Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 162–170, New York, NY, USA. Association for Computing Machinery.

Excell, E. and Al Moubayed, N. (2021). Towards equal gender representation in the annotations of toxic language detection. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 55–65, Online. Association for Computational Linguistics.

Felkner, V., Thompson, J., and May, J. (2024). GPT is not an annotator: The necessity of human annotation in fairness benchmark construction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14104–14115.

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., and Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, 12(1): 8.

Fleisig, E., Abebe, R., and Klein, D. (2023). When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Fleisig, E., Blodgett, S. L., Klein, D., and Talat, Z. (2024). The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

Fort, K. (2016). *Collaborative annotation for reliable natural language processing: Technical and sociological aspects*. John Wiley & Sons.

Fort, K. and Claveau, V. (2012). Annotating football matches: Influence of the source medium on manual annotation. In *LREC - Eight International Conference on Language Resources and Evaluation*, pages 2567–2572, Istanbul, Turkey.

Fort, K. and Sagot, B. (2010). Influence of pre-annotation on POS-tagged corpus development. In *The Fourth ACL Linguistic Annotation Workshop*, pages 56–63, Uppsala, Sweden.

References

Fort, K., Guillaume, B., Constant, M., Lefèbvre, N., and Pilatte, Y.-A. (2018). "Fingers in the nose": Evaluating speakers' identification of multi-word expressions using a slightly gamified crowdsourcing platform. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 207–213, Santa Fe, United States.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of Twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Frenda, S., Abercrombie, G., Basile, V., Pedrani, A., Panizzon, R., Cignarella, A. T., Marco, C., and Bernardi, D. (2024). Perspectivist approaches to natural language processing: A survey. *Language Resources and Evaluation*, 59(2): 1719–1746.

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)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

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

Glickman, M. and Sharot, T. (2024). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 9(2): 345–359.

Gligorić, K., Zrnic, T., Lee, C., Candès, E. J., and Jurafsky, D. (2024). Can unconfident LLM annotations be used for confident conclusions? arXiv preprint arXiv:2408.15204.

Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., Nguyen, L. H., Hao, X., Jaber, B., Reddy, S., Kartha, R., et al. (2023). LLMs accelerate annotation for medical information extraction. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 82–100. PMLR.

Goh, D. H. and Lee, C. S. (2011). Perceptions, quality and motivational needs in image tagging human computation games. *Journal of Information Science*, 37(5): 515–531.

Groves, R. M. and Heeringa, S. G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 169(3): 439–457.

Gruber, C., Hechinger, K., Aßenmacher, M., Kauermann, G., and Plank, B. (2024). More labels or cases? Assessing label variation in natural language inference. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 22–32, Malta. Association for Computational Linguistics.

Gruber, C., Schenk, P. O., Schierholz, M., Kreuter, F., and Kauermann, G. (2023). Sources of uncertainty in machine learning – A statisticians' view. arXiv preprint arXiv:2305.16703.

Guillaume, B., Fort, K., and Lefèbvre, N. (2016). Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3041–3052, Osaka, Japan. The COLING 2016 Organizing Committee.

Haensch, A.-C., Weiß, B., Steins, P., Chyrva, P., and Bitz, K. (2022). The semi-automatic classification of an open-ended question on panel survey motivation and its application in attrition analysis. *Frontiers in Big Data*, 5: 880554.

Harris, C., Aarts, H., Fiedler, K., and Custers, R. (2023). Missing out by pursuing rewarding outcomes: Why initial biases can lead to persistent suboptimal choices. *Motivation Science*, 9(4): 288–297.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition. Springer Series in Statistics. Springer.

He, J., van Ossenbruggen, J., and de Vries, A. P. (2013). Do you need experts in the crowd? A case study in image annotation for marine biology. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pages 57–60.

Heerkens, H., Norde, C., and van der Heijden, B. (2011). Importance assessment of decision attributes: A qualitative study comparing experts and laypersons. *Management Decision*, 49(5): 748–761.

Heim, E., Roß, T., Seitel, A., März, K., Stieltjes, B., Eisenmann, M., Lebert, J., Metzger, J., and Sommer, G. (2018). Large-scale medical image annotation with crowd-powered algorithms. *Journal of Medical Imaging*, 5(03): 034002.

von der Heyde, L., Haensch, A.-C., and Wenz, A. (2024). Vox populi, vox AI? Using large language models to estimate german vote choice. *Social Science Computer Review*.

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*, WWW '15, pages 419–429, Republic and Canton of Geneva, CHE.

Hu, T. and Collier, N. (2024). Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Huang, F., Kwak, H., and An, J. (2023). Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, pages 294–297.

Hube, C., Fetahu, B., and Gadiraju, U. (2019). Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, New York, NY, USA. Association for Computing Machinery.

Jones-Jang, S. M. and Park, Y. J. (2022). How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication*, 28(1): 1–8.

Kern, C., Eckman, S., Beck, J., Chew, R., Ma, B., and Kreuter, F. (2023). Annotation sensitivity: Training data collection methods affect model performance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14874–14886, Singapore. Association for Computational Linguistics.

Keusch, F. (2015). Why do people participate in web surveys? Applying survey participation theory to internet survey data collection. *Management Review Quarterly*, 65(3): 183–216.

Khetan, A., Lipton, Z. C., and Anandkumar, A. (2018). Learning from noisy singly-labeled data. arXiv preprint arXiv:1712.04577.

Kirk, H. R., Bean, A. M., Vidgen, B., Röttger, P., and Hale, S. A. (2023). The past, present and better future of feedback learning in large language models for subjective human preferences and values. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2409–2430, Singapore. Association for Computational Linguistics.

Kreuter, F., McCulloch, S., Presser, S., and Tourangeau, R. (2011). The effects of asking filter questions in interleafed versus grouped format. *Sociological Methods & Research*, 40(1): 88–104.

Kutlu, M., McDonnell, T., Elsayed, T., and Lease, M. (2020). Annotator rationales for labeling tasks in crowdsourcing. *Journal of Artificial Intelligence Research*, 69: 143–189.

Kuzman, T., Mozetič, I., and Ljubešić, N. (2023). ChatGPT: Beginning of an end of manual linguistic data annotation? Use case of automatic genre identification. arXiv preprint arXiv:2303.03953.

Larimore, S., Kennedy, I., Haskett, B., and Arseniev-Koehler, A. (2021). Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Li, J. (2024). A comparative study on annotation quality of crowdsourcing and LLM via label aggregation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6525–6529. IEEE.

Li, M., Shi, T., Ziems, C., Kan, M.-Y., Chen, N. F., Liu, Z., and Yang, D. (2023). CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1487–1505.

Lingren, T., Deléger, L., Molnár, K., Zhai, H., Meinzen-Derr, J., Kaiser, M., Stoutenborough, L., Li, Q., and Solti, I. (2012). Pre-annotating clinical notes and clinical trial announcements for gold standard corpus development: Evaluating the impact on annotation speed and potential bias. *2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology*, pages 108–108.

Lingren, T., Deleger, L., Molnar, K., Zhai, H., Meinzen-Derr, J., Kaiser, M., Stoutenborough, L., Li, Q., and Solti, I. (2014). 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 : JAMIA*, 21(3): 406–413.

Litman, L., Robinson, J., and Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behavior Research Methods*, 47(2): 519–528.

Liu, Q., Jiang, H., Pan, Z., Han, Q., Peng, Z., and Li, Q. (2024). Biaseye: A bias-aware real-time interactive material screening system for impartial candidate assessment. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 325–343.

Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '13, pages 623–631, New York, NY, USA. Association for Computing Machinery.

Lyberg, L., Eckman, S., Roos, M., and Kreuter, F. (2012). Cognitive aspects of dependent verification in survey operations. In *Proceedings of the Survey Research Methods Section, ASA (2012)*, pages 4310–4315, Alexandria, VA. American Statistical Association.

Maaz, K., Trautwein, U., Gresch, C., Lüdtke, O., and Watermann, R. (2009). Intercoder-Reliabilität bei der Berufscodierung nach der ISCO-88 und Validität des sozioökonomischen Status. *Zeitschrift für Erziehungswissenschaft*, 12(2): 281–301.

Martin, D., Hanrahan, B. V., O'Neill, J., and Gupta, N. (2014). Being a Turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, CSCW '14, pages 224–235, New York, NY, USA. Association for Computing Machinery.

Mekler, E. D., Brühlmann, F., Opwis, K., and Tuch, A. N. (2013). Disassembling gamification: The effects of points and meaning on user motivation and performance. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 1137–1142, New York, NY, USA. Association for Computing Machinery.

Miceli, M., Posada, J., and Yang, T. (2022). Studying up machine learning data: Why talk about bias when we mean power? *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP): 34:1–34:14.

Mikulová, M., Straka, M., Štěpánek, J., Štěpánková, B., and Hajic, J. (2022). Quality and efficiency of manual annotation: Pre-annotation bias. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2909–2918, Marseille, France. European Language Resources Association.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T. R., and Mesirov, J. P. (2003). Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology*, 10(2): 119–142.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2): 175–220.

Nowak, S. and Rüger, S. (2010). How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566.

Nédellec, C., Bessieres, P., Bossy, R. R., Kotoujansky, A., and Manine, A.-P. (2006). Annotation guidelines for machine learning-based named entity recognition in microbiology. In *Proceedings of the Data and Text Mining for Integrative Biology Workshop at ECML/PKDD 2006*, pages 40–54.

Oswald, M. E. and Grosjean, S. (2004). Confirmation bias. In *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*, pages 79–96. Psychology Press.

Pangakis, N., Wolken, S., and Fasching, N. (2023). Automated annotation with generative AI requires validation. arXiv preprint arXiv:2306.00176.

Pangakis, N. and Wolken, S. (2025). Keeping humans in the loop: Human-centered automated annotation with generative AI. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1): 1471–1492.

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

Pyatkin, V., Yung, F., Scholman, M. C. J., Tsarfaty, R., Dagan, I., and Demberg, V. (2023). Design choices for crowdsourcing implicit discourse relations: Revealing the biases introduced by task design. *Transactions of the Association for Computational Linguistics*, 11: 1014–1032.

Rabin, M. and Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *The quarterly journal of economics*, 114(1): 37–82.

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., and Tomsett, R. (2022). Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 6(CSCW1): 1–22.

Rehbein, I., Ruppenhofer, J., and Sporleder, C. (2009). Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation. In *Proceedings of the Third Linguistic Annotation Workshop on - ACL-IJCNLP '09*, pages 19–26, Suntec, Singapore. Association for Computational Linguistics.

Richter, A. N. and Khoshgoftaar, T. M. (2020). Sample size determination for biomedical big data with limited labels. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 9(1): 12.

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*, pages 189–199, Virtual Event USA. ACM.

Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., and Vukovic, M. (2011). An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1): 321–328.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., and Tomlinson, B. (2010). Who are the crowdworkers? Shifting demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on*

Human Factors in Computing Systems, CHI EA '10, pages 2863–2872, New York, NY, USA. Association for Computing Machinery.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., and Smith, N. A. (2022). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Schonlau, M. and Toepoel, V. (2015). Straightlining in web survey panels over time. *Survey Research Methods*, 9(2): 125–137.

Settles, B. (2009). Active Learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Shaw, A. D., Horton, J. J., and Chen, D. L. (2011). Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, CSCW '11, pages 275–284, New York, NY, USA. Association for Computing Machinery.

Sheng, V. S., Provost, F., and Ipeirotis, P. G. (2008). Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, pages 614–622, Las Vegas, Nevada, USA. ACM Press.

Singer, E. (2011). Toward a benefit-cost theory of survey participation: Evidence, further tests, and implications. *Journal of Official Statistics*, 27(2): 379–392.

Singh, K., Nakkeerar, R., Sarma, V. V. L. N., Kumar, M., and Office of the Registrar General & Census Commissioner, G. o. I., Ministry of Home Affairs. (2022). Language Atlas of India, 2011. Accessed via Rural India Online (June 17th, 2025).

Suri, S., Goldstein, D. G., and Mason, W. A. (2011). Honesty in an online labor market. In *Proceedings of the 11th AAAI Conference on Human Computation*, AAAIWS'11-11, page 61–66. AAAI Press.

Thorn Jakobsen, T. S., Barrett, M., Søgaard, A., and Lassen, D. (2022). The sensitivity of annotator bias to task definitions in argument mining. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 44–61, Marseille, France. European Language Resources Association.

Toney-Wails, A., Schoeberl, C., and Dunham, J. (2024). AI on AI: Exploring the utility of GPT as an expert annotator of AI publications. arXiv preprint arXiv:2403.09097.

Tourangeau, R., Kreuter, F., and Eckman, S. (2012). Motivated underreporting in screening interviews. *Public Opinion Quarterly*, 76(3): 453–469.

Törnberg, P. (2023). ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. arXiv preprint arXiv:2304.06588.

Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, 14(11): e0224365.

Vădineanu, Ş., Pelt, D. M., Dzyubachyk, O., and Batenburg, K. J. (2023). Reducing manual annotation costs for cell segmentation by upgrading low-quality annotations. In *Workshop on Medical Image Learning with Limited and Noisy Data*, pages 3–13. Springer.

Vaughan, J. W. (2018). Making better use of the crowd: How crowdsourcing can advance machine learning research. *Journal of Machine Learning Research*, 18(193): 1–46.

van der Wal, D., Jhun, I., Laklouk, I., Nirschl, J., Richer, L., Rojansky, R., Theparee, T., Wheeler, J., Sander, J., Feng, F., et al. (2021). Biological data annotation via a human-augmenting AI-based labeling system. *NPJ digital medicine*, 4(1): 145.

Wang, A., Hoang, C. D., and Kan, M.-Y. (2013). Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, 47(1): 9–31.

Wang, P. and Vasconcelos, N. (2023). Towards professional level crowd annotation of expert domain data. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3166–3175, Vancouver, BC, Canada. IEEE.

Wang, X., Kim, H., Rahman, S., Mitra, K., and Miao, Z. (2024). Human-LLM collaborative annotation through effective verification of LLM labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Wu, H., Corney, J., and Grant, M. (2014). Relationship between quality and payment in crowdsourced design. In *Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 499–504. IEEE.

Wyer, N. A. (2010). You never get a second chance to make a first (implicit) impression: The role of elaboration in the formation and revision of implicit impressions. *Social Cognition*, 28(1): 1–19.

Xia, M., Field, A., and Tsvetkov, Y. (2020). Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Yang, Y., Agarwal, O., Tar, C., Wallace, B. C., and Nenkova, A. (2019). Predicting annotation difficulty to improve task routing and model performance for biomedical information extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1471–1480, Minneapolis, Minnesota. Association for Computational Linguistics.

Ybarra, O. (2005). When first impressions don't last: The role of isolation and adaptation processes in the revision of evaluative impressions. *Social Cognition*, 19(5): 491–520.

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

Yin, M., Chen, Y., and Sun, Y.-A. (2013). The effects of performance-contingent financial incentives in online labor markets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 1191–1197.

Yu, D., Li, L., Su, H., and Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apologies. *International Journal of Corpus Linguistics*, 29(4): 534–561.

Zhang, C. and Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, 8(2): 127–135.

Zhang, J., Sheng, V. S., Li, Q., Wu, J., and Wu, X. (2017). Consensus algorithms for biased labeling in crowdsourcing. *Information Sciences*, 382-383: 254–273.

Zhang, Z., Strubell, E., and Hovy, E. (2022). A survey of Active Learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhao, D., Wang, A., and Russakovsky, O. (2021). Understanding and evaluating racial biases in image captioning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14810–14820, Montreal, QC, Canada. IEEE.

Eidesstattliche Versicherung (Affidavit)

(Siehe Promotionsordnung vom 12. Juli 2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist. Bei der Ausarbeitung der Dissertation habe ich die KI-basierten Werkzeuge ChatGPT (OpenAI) und Claude (Anthropic) unterstützend eingesetzt. Die inhaltliche Ausarbeitung sowie die wissenschaftliche Argumentation wurden eigenständig von mir vorgenommen. Für sämtliche Inhalte und deren wissenschaftliche Fundierung übernehme ich die volle Verantwortung.

München, den 10.12.2025 Jacob Beck