# The ‘Invisible’, Often Unhappy Workforce That’s Deciding the Future of AI

Originally published at https://www.unite.ai/the-invisible-often-unhappy-workforce-thats-deciding-the-future-of-ai/

Two new reports, including a paper led by Google Research, express concern that the current trend to rely on a cheap and often disempowered pool of random global gig workers to create ground truth for machine learning systems could have major downstream implications for AI.

Among a range of conclusions, the Google study finds that the crowdworkers’ own biases are likely to become embedded into the AI systems whose ground truths will be based on their responses; that widespread unfair work practices (including in the US) on crowdworking platforms are likely to degrade the quality of responses; and that the ‘consensus’ system (effectively a ‘mini-election’ for some piece of ground truth that will influence downstream AI systems) which currently resolves disputes can actually *throw away* the best and/or most informed responses.

That’s the bad news; the worse news is that pretty much all the remedies are expensive, time-consuming, or both.

## Insecurity, Random Rejection, and Rancor

The first paper, from five Google researchers, is called *Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation*; the second, from two researchers at Syracuse University in New York, is called *The Origin and Value of Disagreement Among Data Labelers: A Case Study of Individual Differences in Hate Speech Annotation*.

The Google paper notes that crowd-workers – whose evaluations often form the defining basis of machine learning systems that may eventually affect our lives – are frequently operating under a range of constraints that may affect the way that they respond to experimental assignments.

For instance, the current policies of Amazon Mechanical Turk allow requesters (those that give out the assignments) to reject an annotator's work without accountability*:

> *'[A] large majority of crowdworkers (94%) have had work that was rejected or for which they were not paid. Yet, requesters retain full rights over the data they receive regardless of whether they accept or reject it; Roberts (2016) describes this system as one that "enables wage theft".*
>
> *'Moreover, rejecting work and withholding pay is painful because rejections are often caused by unclear instructions and the lack of meaningful feedback channels; many crowdworkers report that poor communication negatively affects their work.'*

The authors recommend that researchers who use outsourced services to develop datasets should consider how a crowdworking platform treats its workers. They further note that in the United States, crowdworkers are classified as 'independent contractors', with the work therefore unregulated, and not covered by the minimum wage mandated by the Fair Labor Standards Act.

## Context Matters

The paper also criticizes the use of *ad hoc* global labor for annotation tasks, without consideration of the annotator's background.

Where budget allows, it's common for researchers using AMT and similar crowdwork platforms to give the same task to four annotators, and abide by 'majority rule' on the results.

Contextual experience, the paper argues, is notably under-regarded. For instance, if a task question related to *sexism* is randomly distributed between three agreeing males aged 18-57 and one dissenting female aged 29, the males' verdict wins, except in the relatively rare cases where researchers pay attention to the qualifications of their annotators.

Likewise, if a question on *gang behavior in Chicago* is distributed between a rural US female aged 36, a male Chicago resident aged 42, and two annotators respectively from Bangalore and Denmark, the person likely most affected by the issue (the Chicago male) only holds a quarter share in the outcome, in a standard outsourcing configuration.

The researchers state:

> *'[The] notion of "one truth" in crowdsourcing responses is a myth; disagreement between annotators, which is often viewed as negative, can actually provide a valuable signal. Secondly, since many crowdsourced annotator pools are socio-demographically skewed, there are implications for which populations are represented in datasets as well as which populations face the challenges of [crowdwork].*
>
> *'Accounting for skews in annotator demographics is critical for contextualizing datasets and ensuring responsible downstream use. In short, there is value in acknowledging, and accounting for, worker's socio-cultural background — both from the perspective of data quality and societal impact.'*

## No ‘Neutral’ Opinions on Hot Topics

Even where the opinions of four annotators are not skewed, either demographically or by some other metric, the Google paper expresses concern that researchers are not accounting for the life experiences or philosophical disposition of annotators:

> *‘While some tasks tend to pose objective questions with a correct answer* (is there a human face in an image?*), oftentimes datasets aim to capture judgement on relatively subjective tasks with no universally correct answer (*is this piece of text offensive?*). It is important to be intentional about whether to lean on annotators’ subjective judgements.’*

Regarding its specific ambit to address problems in labeling hate speech, the Syracuse paper notes that more categorical questions such as *Is there a cat in this photograph?* are notably different from asking a crowdworker whether a phrase is ‘toxic’:

> *‘Taking into account the messiness of social reality, people’s perceptions of toxicity vary substantially. Their labels of toxic content are based on their own perceptions.’*

Finding that personality and age have a ‘substantial influence’ on the dimensional labeling of hate speech, the Syracuse researchers conclude:

> *‘These findings suggest that efforts to obtain annotation consistency among labelers with different backgrounds and personalities for hate speech may never fully succeed.’*

## The Judge May Be Biased Too

This lack of objectivity is likely to iterate upwards as well, according to the Syracuse paper, which argues that the manual intervention (or automated policy, also decided by a human) which determines the ‘winner’ of consensus votes should also be subject to scrutiny.

Likening the process to forum moderation, the authors state*:

> *‘[A] community’s moderators can decide the destiny of both posts and users in their community by promoting or hiding posts, as well as honoring, shaming, or banning the users. Moderators’ decisions influence the content delivered to community members and audiences and by extension also influence the community’s experience of the discussion.*
>
> *‘Assuming that a human moderator is a community member who has demographic homogeneity with other community members, it seems possible that the mental schema they use to evaluate content will match those of other community members.’*

This gives some clue to why the Syracuse researchers have come to such a despondent conclusion regarding the future of hate speech annotation; the implication is that policies and judgement-calls on dissenting crowdwork opinions cannot just be randomly applied according to ‘acceptable’ principles that are not enshrined anywhere (or not reducible to an applicable schema, even if they do exist).

The people who make the decisions (the crowdworkers) are biased, and would be useless for such tasks if they were *not* biased, since the task is to provide a value judgement; the people who adjudicate on disputes in crowdwork results are also making value judgements in setting policies for disputes.

There may be hundreds of policies in just one hate speech detection framework, and unless each and every one is taken all the way back to the Supreme Court, where can 'authoritative' consensus originate?

The Google researchers suggest that *'[the] disagreements between annotators may embed valuable nuances about the task'*. The paper proposes the use of metadata in datasets that reflects and contextualizes disputes.

However, it is difficult to see how such a context-specific layer of data could ever lead to like-on-like metrics, adapt to the demands of established standard tests, or support *any* definitive results – except in the unrealistic scenario of adopting the same group of researchers across subsequent work.

## Curating the Annotator Pool

All of this assumes that there is even budget in a research project for multiple annotations that would lead to a consensus vote. In many cases, researchers attempt to 'curate' the outsourced annotation pool more cheaply by specifying traits that the workers should have, such as geographical location, gender, or other cultural factors, trading plurality for specificity.

The Google paper contends that the way forward from these challenges could be by establishing extended communications frameworks with annotators, similar to the minimal communications that the Uber app facilitates between a driver and a rider.

Such careful consideration of annotators would, naturally, be an obstacle to hyperscale annotation outsourcing, resulting either in more limited and low-volume datasets that have a better rationale for their results, or a 'rushed' evaluation of the annotators involved, obtaining limited details about them, and characterizing them as 'fit for task' based on too little information.

That's if the annotators are being honest.

## The 'People Pleasers' in outsourced dataset labeling

With an available workforce that's underpaid, under severe competition for available assignments, and depressed by scant career prospects, annotators are motivated to quickly provide the 'right' answer and move on to the next mini-assignment.

If the 'right answer' is anything more complicated than *Has cat/No cat*, the Syracuse paper contends that the worker is likely to attempt to deduce an 'acceptable' answer based on the content and context of the question*:

> *'Both the proliferation of alternative conceptualizations and the widespread use of simplistic annotation methods are arguably hindering the progress of research on online hate speech. For example, Ross, et al. found that showing Twitter's definition of hateful conduct to annotators caused them to partially align their own opinions with the definition. This realignment resulted in very low interrater reliability of the annotations.'*

* *My conversion of the paper's inline citations to hyperlinks.*

**Published 13th December 2021** – *Updated 18th December 2021: Tags added*

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

**Discover More**