Towards fair and pro-social employment of digital pieceworkers for sourcing machine learning training data

Annabel Rothschild
Georgia Institute of Technology
arothschild@gatech.edu

Justin Booker*
DataWorks, Georgia Institute of Technology

Christa Davoll*
DataWorks, Georgia Institute of Technology

Jessica Hill*
DataWorks, Georgia Institute of Technology

Venise Ivey*
DataWorks, Georgia Institute of Technology

Carl Disalvo
Georgia Institute of Technology

Benjamin Rydal Shapiro
Georgia State University

Betsy Disalvo
Georgia Institute of Technology

ABSTRACT

This work contributes to just and pro-social treatment of digital pieceworkers ("crowd collaborators") by reforming the handling of crowd-sourced labor in academic venues. With the rise in automation, crowd collaborators treatment requires special consideration, as the system often dehumanizes crowd collaborators as components of the "crowd" [41]. Building off efforts to (proxy-)unionize crowd workers and facilitate employment protections on digital piecework platforms, we focus on employers: academic requesters sourcing machine learning (ML) training data. We propose a cover sheet to accompany submission of work that engages crowd collaborators for sourcing (or labeling) ML training data. The guidelines are based on existing calls from worker organizations (e.g., Dynamo [28]); professional data workers in an alternative digital piecework organization; and lived experience as requesters and workers on digital piecework platforms. We seek feedback on the cover sheet from the ACM community.

CCS CONCEPTS

• **Social and professional topics**; • **Professional topics**; • **Computing profession**;

KEYWORDS

Platform labor, crowd working, Amazon Mechanical Turk, computing ethics, crowd collaboration

ACM Reference Format:
Annabel Rothschild, Justin Booker, Christa Davoll, Jessica Hill, Venise Ivey, Carl Disalvo, Benjamin Rydal Shapiro, and Betsy Disalvo. 2022. Towards fair and pro-social employment of digital pieceworkers for sourcing machine learning training data. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts), April 29–May 05, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3491101.3516384

*These authors contributed equally to the work and are listed by alphabetical order of last name.

CHI '22 Extended Abstracts, April 29–May 05, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9156-6/22/04.
https://doi.org/10.1145/3491101.3516384

1 INTRODUCTION AND MOTIVATION

The use of crowd-working platforms (e.g., Amazon Mechanical Turk [AMT], Clickworker [CW], and Microworker [MW]) to build large-scale machine learning training datasets is both widely accepted and prevalent in academic research practice [36]. What is also prevalent is academic researchers abusing the power imbalance of these platforms, leaving digital pieceworkers [2] at the mercy of their academic requesters [37]. Given the dehumanizing title workers are given – they are only a subunit of the larger "crowd" [41] – these practices are not surprising; indeed, while there are bad actors, the distributed nature of the platforms may cause requesters to appear to enact these malpractices out of misunderstanding more than malintent. For this purpose, we propose a cover sheet describing precise hiring and employment practices of academic collaborators (or "crowd collaborators"[1]) engaged through crowd-working platforms. The cover sheet is to be submitted with the publication of the resulting work by researchers in academic venues. The design of the cover sheet is inspired by the Datasets for Datasheets project [13], to be required by academic venues accepting the results of crowd collaborator labor, namely machine learning training datasets. As Hießl argues, crowd collaborators do not have traditional employment contracts to rely on and that a new form of contract must be developed to address the complexity of digital piecework [16]; we present this cover sheet as a first step in that direction. By surfacing this information at the institutional level we hope to 1) inform requesters of the best practices if they are unaware, and 2) certify respectful treatment of crowd collaborators, especially given the calls to substitute digital piecework for jobs lost in the face automation [20, 29].

Our intervention centers academic requesters for two reasons. First, we choose to highlight the role of requesters in these platforms as the power balance is inevitably shifted away from those performing the labor to those providing it, due to the oversupply of

[1]While we cannot find the original use of this term, we are sure it has been used in prior work and the original author(s) should be credited. Our use of the term pulls from the collaborative nature of digital piecework workers and academic requesters who use their services (described in [39, 45, 53]); our intention is to highlight the value of the work that digital pieceworkers perform and highlight their contributions.

labor and undersupply of available tasks [14, 23]. Previous interventions have sought to reform the digital piecework labor system from other angles. One of the most well-known, the TurkOpticon project [40], helps crowd collaborators source much-needed information about the requesters and tasks they encounter. However, TurkOpticon's founders realized it became a permanent, relied upon feature of the ecosystem rather than forcing Amazon's hand to create permanent, platform-implemented safeguards for workers [17]. Similar efforts to improve the platform from the worker side include the Crowd-Worker plugin [7], Crowd Guilds to unite workers [50], and a worker-owned cooperative model alternative [43]. These innovations either beneficially augmented the experience of workers or proposed alternatives; however, they require external funding and, in some cases, forgoing immediate profits for long-term vested interests, which is not an option for workers who need immediate payout [35]. Further, there may exist inequalities in the way different crowd collaborators are rated, where applicable [19]. Similarly, while unions promote higher income and feeling of community between workers [49], these digital work platforms sometimes act against them, as in the case of Amazon Mechanical Turk banning the account(s) of small, collaborative groups [14].

The We Are Dynamo project [32], which unites workers to achieve collective action, has provided among numerous outcomes a best practices guide for academic requestors, however there's no system to enforce academic requestors' adherence to them. Whiting et al. [51] inspired our decision to shift the focus to requesters as they trust workers to report the time they took to complete the task (or a reasonable approximation) in order to guarantee proper payment for work performed. Requesters hold a great deal of power over crowd collaborators, as does the platform. However, the latter seems impervious to improvement in the short term, as described by the operators of TurkOpticon, whose improvements to the AMT platform for workers were not adopted by AMT itself [17]. We hope to combine the common issues surfaced by these efforts and provide a way to operationalize their findings and concerns by mandating compliance at the institutional level, similar to the IRB process for human subjects. Our goal is to extend existing knowledge about what a fair requestor-worker dynamic looks like into a formal reporting system to create a more just and respectful workplace for crowd collaborators.

Second, we highlight the role of requesters from inside the Academy. As requestors, academics and our industry collaborators – as highlighted by Scheuerman et al.'s study of computer vision researchers – are failing to meet basic standards (e.g., clear standards for terms of employment) for fair digital piecework practices [36] despite the popularity of such platforms [18]. Further, as we continue to confront the biases embedded in our research designs and products with regards to data, we must acknowledge that – in many cases – they are the result of our own oversight and overly-generalized practices rather than the fault of our crowd collaborators [1, 8, 25, 26] and that once compiled, datasets have long lives [9, 21]. Along with prescient data about the terms of employment, we ask that requesters engage in a reflection of what values or experiences are reflected in the data work they request.

This paper presents the cover sheet as a specific contribution, but also seeks to engage in dialogue with the larger ACM academic community to evolve the notion of cover sheets and other related ideas. We do not seek to end the practice of sourcing digital piecework through crowd-work platforms. We recognize both the research and employment opportunities that these platforms provide, especially with respect to workers who may have preoccupying care-giving tasks, difficulty travelling to a workplace, or face discrimination in the workplace. Rather, we hope to institute a more sustainable practice that engages crowd workers as collaborators, acknowledging both the injustices that academic requesters have perpetrated on crowd workers and the changing nature of labor in the face of automation.

2 METHODOLOGY

Our methodology for constructing the cover sheet is based on a mix of first-hand experiences in the reflective style of [6] as well as direct observation by professional data workers, and finally previous findings by academic and crowd worker community bodies (e.g., We are Dynamo [32]). The first version of the cover sheet was designed based on the first author's observations from the experiences above and existing literature, structured along the comparison axes the data workers highlighted in Fig. 1. The data workers then provided feedback during a 1-hour session which resulted in the second version of the cover sheet shared in this paper. In the explanation for the different pieces of information, the data workers are quoted directly or summarized in brief from the research records collected during the engagements listed below.

2.1 Setting

The second through sixth authors are employees of DataWorks, a work training program for developing the skills of a mid-skill data worker incubated in the Georgia Tech College of Computing. The program aims to broaden participation in the everyday work of data collection, cleaning, and basic analysis. DataWorks' employees (the "data workers") are people from economically disadvantaged neighborhoods and underrepresented groups in computing and DataWorks aims to assist them in finding solid, middle-class jobs in data work. In many ways, DataWorks functions as an alternative to more classical digital piecework platforms like AMT, CW, and MW. The data workers engage with a variety of client projects, a portion of which resembles classic image recognition or natural language processing sentiment analysis tasks that would be found on the aforementioned platforms, along with more typical spreadsheet-based projects. Unlike those platforms, however, employees of DataWorks are full employees of Georgia Tech and therefore receive a competitive hourly wage ($17.35), health care (USA-specific), other fringe benefits, organization-provided computers and workspace and work a 40-hour week with regular hours.

Further, the data workers have extensive input on client projects and engage in dialogue directly with the client, including – depending on the project – initial training sessions, clarification questions, and project presentation at the conclusion. DataWorks' client projects are longer term than discrete digital piecework tasks; for example, the data workers identified and summarized the events of close to 900 cartoons for a single requestor. The data workers have a skillset that is – with regards to this kind of digital piecework – therefore comparable to an experienced, professional worker on the more classical platforms (e.g., AMT, CW, MC). While the data

workers are not full-time crowd collaborators, their expertise and experience play an important role: they are aware of alternative structures to classical platform work – via DataWorks – and, given the research setting of the workshop, they can take the time to ideate and critique. While investigations that center experienced, professional crowd workers are of immense importance, we believe that adding the accounts of the data workers is an important contribution.

2.2 Construction

Over the course of seven days in the summer of 2021, the data workers, along with the first author, engaged in a reflective workshop to compare the experience of working for DataWorks with that of a crowd collaborator on three digital piecework platforms: AMT, CW, and MW. Other platforms were initially investigated but the workers were not able to complete work on the platforms due to the location requirements for Sama (formerly Samasource), full-time requirements for LeadGenius (formerly MobileWorks [27]), and lack of available tasks for Appen (acquired the former Figure Eight platform). The workshop took place under the auspices of university IRB approval and the data workers were paid their normal hourly wage while engaging in the workshop. The workshops enable us as researchers to better understand work practices and provide the workers with domain-specific skills and business practice.

The workshop was intended to identify what aspects of employment for digital piecework DataWorks was getting right and which aspects the organization could improve. The workshop was open-ended and began during the fourth week of a 10-week summer tutorial course designed and facilitated by the first author on the politics of data and key data cleaning and standardization skills. The point of the workshop was to directly engage with and observe alternative employment systems for digital pieceworkers and compare and contrast experiences on those platforms through discussion. Unlike other experiential work on digital piecework platforms (direct observation by researchers or interviews with, or observations of, crowd collaborators on those platforms), the data workers have the professional experience of being digital pieceworkers and given a lack of time pressure, were able to reflect on their experience and brainstorm alternatives. The data workers' impressions were collected through four kinds of engagements:

- The data workers engaged individually with the platform (to mimic the isolated nature of digital piecework), including signing up and working on the three platforms. The data workers recorded their impressions on shared and individual note-taking documents. Duration = 7 hours, broken into multiple sessions. The data workers kept shared and individual documents of running notes and discussed their experiences with the first author in their regular interviews (see item four of this list).
- The data workers worked communally on a given task, with one operating a computer from which the task was projected on a large screen, and all team members engaged together on the task, more akin to the collaborative setting in which the data workers usually operate. During this session, the data workers trialed the TurkOpticon browser add-on and were introduced to other AMT community resources, such as TurkerView forum. At the end of the session, the data workers compared their experiences on the three platforms (see Fig. 1). Duration = 90 minutes. The session was recorded and the first author took notes.
- The data workers described their experiences to the sixth author, who was unfamiliar with the platforms, in a 90-minute session, intermittently working on tasks together during that time to demonstrate their reflections. Duration = 90 minutes. The sixth author took notes and the data workers expressed their recollection in their regular interviews (discussed below).
- Finally, throughout the workshop, the data workers met individually with the first author to reflect on their experiences in semi-structured interviews. Duration = roughly 15 minutes per worker per week, for three weeks. The first author audio recorded interviews and took notes during the sessions.

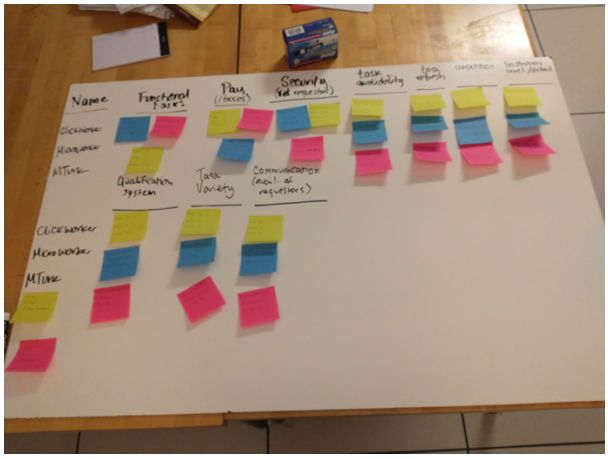

Figure 1: A comparison chart of the three platforms by three of the data workers. As the sticky notes are handwritten, a typed version is available for download at https://annabelrothschild.com/documents/alt-chi-22/Fig1-text-translation.docx

In all, the data workers accrued more than 10 hours each of experience on the three platforms, with the majority on AMT, through two 90-minute tutorial sessions and 7 hours of independent work spread over multiple days. Some workers did not attend all workshop sessions due to other conflicts, but all workers completed at least 10 hours of experience on the platforms combined. Where possible, the data workers never "cashed out" payment. Because they were forced to "cash out" to register for some platforms they earned $8.98, which went towards snacks for the DataWorks office. The data workers were not required to provide their personal information in order to use the digital pieceworker platforms.

2.3 Researcher positionality

Authors 2-6 on this paper are the data workers whose insights and backgrounds significantly contributed to the design and development of the coversheet described in this paper. Additionally, the

first author, who interacted most with the data workers to design and development the coversheet, has extensive experience (spanning four years) as both a worker and requestor on AMT. These experiences include early experiences working at Wellesley College, first with Dr. Eni Mustafaraj then with Dr. Ada Lerner, working on AMT in order to understand how human intelligence tasks (HITs) should be developed and later as requestor for both quantitative and qualitative tasks. Practices developed in the Cred – motivated by [9] – and Security & Privacy labs (respectively) with both PIs and other student researchers inform the provided practical examples of how requesters can, among other things, calibrate payment and employ respectful demographic questionnaires. These experiences likewise significantly influenced the design and development of the coversheet described in this paper, for example, to inform portions of the guidelines that refer specifically to the ways in which requesters can structure and reflect on their digital piecework tasks.

3 COVER SHEET ITEMS

The cover sheet[2] is meant to surface information about the conditions of collaborator employment and build an ecosystem that is more respectful of crowd collaborators. Features of the AMT platform are often used as examples because the platform is one of the more robust and popular with the computing community; however, there are identical features and mechanisms on other platforms. Some of this information (e.g., payment) has often been reported but not uniformly [37]. We do not think filling out the cover sheet should take more than an extra hour for the research team; however, the tasks required (e.g., follow up with accepted or rejected crowd collaborators) may require extra time during the experimental phase of the research. We argue below that each of these additional tasks has a meaningful effect for crowd collaborators and should be required as such. As we describe in 4 (Future directions), we will be confirming this (and iterating on, as necessary) in a field experiment. An additional practice that academic requesters should consider is monitoring their reputation (via their profile) on forums like TurkerView[3] to proactively catch problems crowd collaborators encounter, a practice employed by the Cred Lab.

We build off the recommendations of the Dynamo guidelines [54] and use the ± symbol to indicate reiteration and expansion of recommendations developed by the We Are Dynamo movement.

1. **Basic information**. This information should be described to help the academic community receiving the contributions of the crowd collaborators assess the context of the task.
   a. *The platform used (e.g., AMT).* For reproducibility – described further in [31] – and to assess per platform specific features, some of which are discussed in [4].
   b. *The requestor name used to post the task* ±. Providing clear, factual information in the requestor name (e.g., Prof. X, University Y Lab Z) can help crowd collaborators understand who they are working for and track the progress of individual tasks in the post-submission phase. The former allows crowd collaborators to help describe the nature of their work and recognizes them as collaborators, while the latter helps collaborators track their salary and address concerns. As a data worker stated with regards to platform work, "*Can you even use this...can you put it on your resume, is it respected work*?" Acknowledging the skilled labor required to complete HITs for ML training data, requesters should allow their collaborators to signal their expertise. In addition, if an academic requestor has a faulty task and fails to state (or misstates) follow up information, the researcher can be found via their public institutional profile online.
   c. *The full HIT name and short description with task category* ±. More information readily available to crowd collaborators allows them to cut down on the significant labor of sifting through available work [7, 35, 51]. The data workers also highlighted the importance of knowing the task category (e.g., image "tagging" for recognition) in conventional crowd collaborator language (e.g., "chat with a bot" for NLP conversational work).
   d. *Contact information given to the crowd collaborators and designated team member(s) who monitored inbox* ±. The contact information (e.g., email) provided to crowd collaborators should be made available to ensure that it is accessible. Designated research team member(s) should be "on call" to monitor the contact inbox to ensure that crowd collaborators can receive follow up within a reasonable time frame; the data workers' consensus was 24 hours was appropriate and this number should be confirmed in future work. For example, in Drs. Mustafaraj and Lerner's labs, HITs were posted with a contact email address that would automatically forward to the PI and research assistants on the project, or a lab address that research assistants running studies could access and the PI could review; the individual student(s) running the experiment are then responsible for monitoring that email address. Particularly when apprenticing researchers are involved (i.e., students) who may be new to running experiments on crowd labor platforms, a more experienced member of the team can ensure that the apprentices are engaging properly with crowd collaborator inquiries.
   e. *IRB consent form, if applicable* ±. For archival purposes; can be attached to the cover sheet as supplemental material to ensure coherence with.
   f. *Warnings provided about potentially sensitive activities or topics.* Crowd collaborators should be given enough information about sensitive topics in a task so they can make an informed decision about accepting the HIT without having to scroll through multiple warning screens – or worse – be forced to abandon the HIT partway. This respects their time and does not affect their return rate, which can be used as a collaborator qualification on AMT, for example.
   g. *Time(s) of day and day(s) of week HIT posted, including the number of HITs posted in (each/the) batch.* This information should be provided to gauge potential population bias or impacts on crowd collaborator lives. For example, requesters should respect local time zones – and, if hoping to

[2] An example of the cover sheet as a fillable PDF, along with a completed example cover sheet, are available at: https://annabelrothschild.com/projects/alt.chi-22/prosocial_crowd_collaborator_recruitment_guidelines
[3] https://turkerview.com/

achieve a global reach, should post their tasks at times that are conducive to collaborators in those locations [14, 20].

2. **Crowd collaborator treatment**. This section addresses concerns that directly relate to the treatment of employed crowd collaborators, including – but not limited to – fair compensation, ensuring a right to privacy and security, and structuring a HIT as accessible as possible.
   a. *Terms of employment.*
      i. *Number of collaborators desired and proposed payment per crowd collaborators, along with any later bonuses paid out.* The number of collaborators refers to the number of distinct individuals who complete the tasks to determine the diversity of the collaborator population who performed a certain HIT, which may have implications for the use of a given dataset, given the subjectivity of the work. Further, requesters should state how much they intended to pay collaborators per HIT and how they arrived at that number as described in [30]. Suggested methods include interacting with the locality-respecting calculator built by Sinders [42] and the one-line of code used to guarantee a $15/hr wage[4] built by Whiting et al. [51]. Additional bonuses should be described; e.g. via Whiting et al.'s mechanism or as part of total compensation used to follow up with individual crowd collaborators (AMT, for example, requires a minimum 1 cent USD for the "bonus" payment mechanism which can be used to communicate with collaborators after they have finished a HIT). The Dynamo guidelines for academic requesters state that $0.10 USD per minute is considered an effective pay floor and that "tasks paying less than $0.10 a minute are likely to tap into a highly vulnerable work pool and constitutes coercion." [54] While many requesters operating out of the United States may consider applying their state or district minimum wage, consider there are a wide range of minimum wages in the US (from $7.25 to $15 at the time of writing). Without asking the collaborator about their location (with IP address being a poor proxy as Whiting et al. describe [51]) it is difficult to ascertain proper minimum wage; for this reason, Whiting et al 2019 default to $15 per hour. Further, d'Eon et al. describe the mutual benefice of fair wages and how they might be calibrated [11].
      ii. *Number of crowd collaborators accepted and percent accepted rate; number of crowd collaborators rejected and percent rejection rate.* A high rejection rate (context specific, but generally more than 10%) can indicate a faculty HIT; for example, a mechanical issue or a lack of clear instructions. Further, rejection without clear rationale can indicate that collaborators were unfairly rejected so the requestor could get more labor for less compensation. A high rejection rate should be explained and follow up action should be described, such as a soft-reject (compensate collaborators who did the task correctly to their understanding but incorrectly for the purposes of the researcher; in this case, the researcher "accepts" the HIT and pays for the work, given that it was their mistake). For clear-cut tasks, rejection rates are expected to be low. The data workers described engaging in a task that required copying and pasting the results of a Google search query ranking about which there was no ambiguity; they were rejected either without rationale or a confusing "nice work!" message which indicated either malintent by the requestor or accidental action (the requestor never responded to follow up messages from the data workers).
      iii. *Criteria for rejection (list)* ±. Pursuant to immediately preceding item, requesters should summarize criteria for rejection after reviewing multiple rejection-worthy entries and follow up with individual crowd collaborators to communicate cause for rejection. This assures the collaborator that the rejection was legitimate (e.g., a spam entry or failing reasonable "attention checks") if there was a mistake on the part of the requestor (seeming accidental rejection) provides the collaborator with a clear way to request clarification. Where possible, rejection rationale should be communicated as "fruitful feedback" [28].
      iv. *Follow-up method to communicate rejection/acceptance for each crowd collaborator.* Pursuant to the preceding two items, crowd collaborators often site lack of reasonable follow-up and communication from requesters are a major problem [5, 49]. All follow up should include the HIT name and requestor name in communication. Before posting the task, researchers should assess the follow-up mechanisms of the platform and if they must collect additional information to engage in follow-up add that to their task with clear rationale for doing so and allow given collaborators the opportunity to opt-out (in case they do not want to provide a mechanism for follow-up out of privacy concerns). Any future contact information collected must be allowable by the platform Terms of Service.
      v. *If disallowing multiple submissions by a given crowd collaborators, state mechanism used to do so* ±. A platform's "blocking" feature should not be used as it limits the future work available to a crowd collaborator by disbarring them from future, unrelated tasks from the same requestor. Instead, on AMT for example, requesters should make use of the "qualification" mechanism to disbar multiple entries for a single HIT. If a qualification mechanism is used, make the purpose of qualification the qualification name (e.g., "July2021StudyNoMultipleSubmission") to help crowd collaborators track HITs completed and reduce ambiguity around random qualifications [15].
      vi. *List any required collaborator qualifications and rationale for them.* Extensive use of qualifications limits both the pool of available crowd collaborators and the work available to crowd collaborators. Previous research has shown that not all distinctions are necessarily reliable

[4] As of 11/22/2021, $15 USD in the following highly populous countries = Chinese Yuan: 95.79; Indian Rupee: 1116.05; Indonesian Rupiah: 213745.50; Pakistan Rupee: 2628.63; Brazilian real: 83.78; Nigerian Naira: 6172.72

metrics [23]. Given the over-subscription of workers compared to the number of HITs available [16], requesters should be conscientious to extend work to all legitimately qualified collaborators, pursuant to task type characteristics. Given the prevalence of unfair rejections, requesters should be sensitive to using approved HITs as a qualification metric; collaborators who are unfairly rejected must then work a high number of HITs correctly to fix their acceptance ratio, which can force them into low-paying and exploitative work.

vii. *State any pre-test tasks.* Sometimes HITs require particular qualifications that must be described by the individual collaborator. If requesters are seeking specific demographic characteristics, for example, they should consider using a platform that caters more directly to that need – for example, Prolific appears to be one such, but individual requesters should confirm this. If collaborators will potentially be disbarred from a task given their pre-test results, they should still be compensated for their time as they produced labor and information (if incorrect) for the requestor. Malicious requesters may require extensive pre-test information that allows them to get the majority of their HIT done despite rejecting most (or all) crowd collaborators. Rejected crowd collaborators are then not compensated despite effectively completing the HIT; this reporting provides one mechanism to eradicate that behavior.

viii. *State average payout speed for HIT(s).* Simply because their labor occurs in a distributed fashion does not mean crowd collaborators are less deserving of a regular, predictable paycheck. Academic requesters should make extensive effort to review submissions within 24 hours and release payment at that time. Given the varying speeds it takes the platforms to transfer that compensation to the collaborator, this helps collaborators estimate their future earnings with better accuracy.

b. *Privacy and security.*

i. *State technical format of task; e.g., were collaborators required to open a new browser window (distinct from the HIT page on the platform's website) or download any additional software? State all format(s) and rationale for each.* Previous work [33, 34] has demonstrated that working on crowd work platforms introduces an individual to a number of cybersecurity and privacy concerns. Where possible, the activity for a HIT should be contained in the official HIT page on the crowd work platform. If additional screens or software are necessary, crowd collaborators should be informed why those steps are necessary and how they will appear to reduce surprise and give crowd collaborators a chance to consider whether or not they feel comfortable engaging in the HIT. There are also considerations raised by [12] about the limitations of HCI work on piecework platforms which should be considered. One of the data workers described their initial impressions of HITs on one platform: "some of them are kind of weird" in reference to a posting that asked the worker to upload selfies and another that asked for a copy of the worker's government ID. "Think it's kinda sketch but I'll do it," another worker said of tasks that required them to open new browser windows to a provided link.

ii. *If collecting user demographics, was a "prefer not to answer" option for all questions available and were collaborators clearly informed that they would not be penalized for selecting that option?* Further: were collected demographics protected by a privacy protocol and was this protocol made available should collaborators want to see it? One benefit of micro task platform sites is that they often allow a crowd collaborator to remain anonymous to the requestor. This may allow some individuals to gain an income where they might otherwise be unable to engage in work for fear of discrimination, persecution, or ridicule. While collecting demographic information may be important (2.4.1), care should be taken to allow individuals to protect some (or all) of their demographic information. Further, as many tools to help automate collection of HIT responses automatically record possibly identifying information (e.g., location and/or IP address), special care should be taken to protect potentially identifying demographic information, even when the HIT activity does not require sensitive information. Individual collaborators may not want to identify as being such for any variety of reasons and careful care should be taken to prevent them from being deanonymized, even if the likelihood is extremely low. For example, [47] illustrates the need for privacy for low-income women in the Global South.

c. *HIT structure and format.*

i. *(A) Average satisfactory completion time in trial runs; (B) trial population and size of population task piloted with; and (C) approximate relationship of population to crowd collaborators.* HITs should be tested for both functionality and estimated time to complete. Requesters may try to determine this information but are often incorrect [51]. In part, the validity of the approximation of the pilot population to the crowd collaborator population may be difficult to ascertain. Chapter 3 of [3] provides a starting point for comparing key demographic factors of crowd collaborators compared to the pilot population and can inform assumptions about approximation validity. In Drs. Mustafaraj and Lerner's labs, student researchers working on different projects pilot each other's studies; however, given the topicality of each lab and that student researchers generally have high literacy as college students who have been trained in such, additional time is added to compensate for crowd collaborators who may not have had the same opportunities or have the same general familiarity with the topic or task type.

ii. *(A) Range and median of completion times for accepted crowd collaborators; (B) range and median of completion times for rejected submissions by crowd collaborators.* Requesters should pay attention to the amount of time required to complete their HIT. Deploying HITs

in small batches provides one way to ensure that completion times and compensation for such are fair – if this ratio is not reasonable, future batches can increase compensation for the task and past collaborators should be compensated additionally via bonuses as appropriate. Requesters should also be aware of the polychronicity – or multitasking habits – of crowd collaborators on some platforms [22] and put the HIT upper limit at a generous time allotment.

iii. *Confirm use of persistent progress bar or other indicator of progress.* The use of a progress indicator allows crowd collaborators to determine time spent on the task so far (relative to compensation) and helps them make an informed choice about whether or not to continue with the task. This should not be an issue in HITs that have calibrated payments to time spent with accuracy.

d. *Data collected.*

i. *Describe any steps taken to root out automated responses or malicious entries.* CAPTCHAs and "attention check questions" (often simple calculations, e.g., "what is two plus three?", or hidden directions, e.g., "regardless, check the fourth option below" after a long block of question text) help requesters root out automated or insincere entries [24]. However, requesters should ensure that their methods are accessible to collaborators who have hearing or visual impairments and may be using alternate technologies. Estimated time to complete these authenticity/sincerity checks should be compensated and payment should consider the time it may take a collaborator using assistive technology to complete. The Dynamo guidelines also suggest double checking functionality of all attention check devices [54].

ii. *Whether or not crowd collaborator demographics were collected; rationale for choice; and basis for demographic categories (if used).* There are a variety of reasons for which requesters may or may not choose to collect particular pieces of collaborator demographics. In some cases, particular demographic experiences may be correlated with cognitive biases that affect how the ensuing dataset should be understood [10]. In other cases, collecting demographic information may require extra labor from crowd collaborators, which can be frustrating when the cause for collection is not clear [52]. One of the data workers described situations in which they felt their demographic background (as it shaped their experience) was relevant, citing image recognition in a case where they felt it mattered depending on the kind of image being labeled. In contrast, if they were providing textual translation of a photographed word, they felt it was less important. In cases where demographics were requested, one data worker suggested that the requesters should share their own demographic background, to help contextualize the work and help the collaborator gauge the motivation of the request, which all the other data workers present agreed was important. If demographics are collected, the language used to request that information should be carefully considered and respectful of the diversity and variety of human experience and background. For example, Scheuerman et al. demonstrate the reductive language used in computing around gender that does not reflect the diversity of gender in the human population at large [38]. Free-text options and multiple-selection checkboxes may facilitate this, along with an opt-out choice for all questions, as described above (2.2.2).

4 FUTURE DIRECTIONS

We view the proposed cover sheet as a "living" document: we hope to accrue feedback for the contents of the cover sheet through the SIGCHI community, to present a version for practical use that covers as many concerns about just labor practices as possible. We will post these preliminary guidelines on GitHub and Google Docs (& Forms) to accrue feedback[5]. We particularly seek just practices for crowd collaborators operating outside of the United States or who are undocumented in the US, given that our experience is both US and documented-resident centric. For example, currency and payout format may be concerns we should investigate more deeply when the default on some platforms (e.g., AMT) is documented US residents operating in the US. We also hope for suggestions about accessible formatting of HITs, such as those introduced by [46] and accessible platforms such as BSpeak [48]. Finally, we seek to engage with professional digital pieceworkers to compile their feedback and will investigate respectful, collaborative ways to engage with that community.

Following a feedback cycle, we plan to conduct in situ experiments with academic researchers utilizing crowd work platforms for ML data work to see how the cover sheet affects their work, both in how they deploy their tasks and how they later use the data collected. We will then explore how to institutionalize the practice of mandatory reporting crowd collaborator employment terms in venues where such work is presented – for example, ML community conferences and gatherings.

We believe that the push towards automation and the ML training dataset development that requires an immediacy of action to ensuring proper behavior by academic requestors. While many academic requesters using crowd platform labor for this purpose may be interacting with crowd collaborators with sincerity and best intentions, it is still necessary to push for institutional norms that guarantee just treatment of crowd collaborators. There are also other institutions – for example, individual Institutional Review Board (IRB) programs in the United States – and funding bodies, as well as professional associations (such as the Association for Computing Machinery), who should be considered as sites of enforcement. Along with supporting high-level pushes, like that of the European Trade Union Confederation [44], we hope to provide immediate improvement in the conditions of workers on crowd labor platforms, particularly those used by academic researchers for ML data work.

[5]Links collected here: https://annabelrothschild.com//projects/alt.chi-22/pro-social_crowd_collaborator_recruitment_guidelines

5 CONCLUSION

In this work we propose a cover sheet to accompany the submission of projects to academic venues that require the labor of crowd collaborators. Our goal is to surface the conditions in which crowd collaborators are operating and ensure that academic requesters – specifically those seeking ML training data – treat crowd collaborators fairly and respectfully. Through alt.chi we seek feedback on the first iteration of the cover sheet and hope to discuss aspects of crowd collaboration terms which we may be overlooking, such as concerns of crowd collaborators located outside of the United States, along with those that utilize assistive technologies to work on crowd labor platforms.

ACKNOWLEDGMENTS

This work was enriched by conversations with Drs. Eni Mustafaraj and Ujwal Gadiraju. Preliminary ideas for the cover sheet were presented at The Global Labors of AI and Data Intensive Systems workshop at CSCW '21; we thank the organizers and attendees for their feedback. This work was funded by The National Science Foundation Grant no: GR0006226 along with generous support provided by the Georgia Tech College of Computing and the Constellations Center for Equity in Computing at Georgia Tech.

REFERENCES

[1] Ali Alkhatib. 2021. To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3411764.3445740

[2] Ali Alkhatib, Michael S. Bernstein, and Margaret Levi. 2017. Examining Crowd Work and Gig Work Through The Historical Lens of Piecework. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 4599–4616. https://doi.org/10.1145/3025453.3025974

[3] Daniel Archambault, Helen Purchase, and Tobias Hoßfeld (eds.). 2017. Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments: Dagstuhl Seminar 15481, Dagstuhl Castle, Germany, November 22 – 27, 2015, Revised Contributions. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-66435-4

[4] Frank Bentley, Kathleen O'Neill, Katie Quehl, and Danielle Lottridge. 2020. Exploring the Quality, Efficiency, and Representative Nature of Responses Across Multiple Survey Panels. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3313831.3376671

[5] Alice M. Brawley and Cynthia L.S. Pury. 2016. Work experiences on MTurk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior* 54: 531–546. https://doi.org/10.1016/j.chb.2015.08.031

[6] Matthias Budde, Anja Exler, Till Riedel, Micheal Beigl, and Andrea Schankin. 2017. Lessons from failures in designing and conducting experimental studies: a brief anecdotal tutorial. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, 992–999. https://doi.org/10.1145/3123024.3124392

[7] Chris Callison-Burch. 2014. Crowd-Workers: Aggregating Information Across Turkers to Help Them Find Higher Paying Work. In *Proceedings of the Seconf AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2014, November 2-4, 2014, Pittsburgh, Pennsylvania, USA*. Retrieved from http://www.aaai.org/ocs/index.php/HCOMP/HCOMP14/paper/view/9067

[8] Evgenia Christoforou, Pinar Barlas, and Jahna Otterbacher. 2021. It's About Time: A View of Crowdsourced Data Before and During the Pandemic. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3411764.3445317

[9] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society* 8, 2: 205395172110359. https://doi.org/10.1177/20539517211035955

[10] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A Checklist to Combat Cognitive Biases in Crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9, 1: 48–59.

[11] Greg d'Eon, Joslin Goh, Kate Larson, and Edith Law. 2019. Paying Crowd Workers for Collaborative Work. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW: 1–24. https://doi.org/10.1145/3359227

[12] Leah Findlater, Joan Zhang, Jon E. Froehlich, and Karyn Moffatt. 2017. Differences in Crowdsourced vs. Lab-based Mobile and Desktop Input Performance Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 6813–6824. https://doi.org/10.1145/3025453.3025820

[13] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. *arXiv:1803.09010 [cs]*. Retrieved December 11, 2020 from http://arxiv.org/abs/1803.09010

[14] Mary L. Gray and Siddharth Suri. 2019. *Ghost work: how to stop Silicon Valley from building a new global underclass*. Houghton Mifflin Harcourt, Boston.

[15] Benjamin V. Hanrahan, David Martin, Jutta Willamowski, and John M. Carroll. 2018. Investigating the Amazon Mechanical Turk Market Through Tool Design. *Computer Supported Cooperative Work (CSCW)* 27, 3–6: 1255–1274. https://doi.org/10.1007/s10606-018-9312-6

[16] Christina Hießl. 2018. Labour law for TOS and HITs? reflections on the potential for applying 'labour law analogies' to crowdworkers, focusing on employee representation. *Work Organisation, Labour & Globalisation* 12, 2: 38. https://doi.org/10.13169/workorgalaboglob.12.2.0038

[17] Lilly C. Irani and M. Six Silberman. 2016. Stories We Tell About Labor: Turkopticon and the Trouble with "Design." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4573–4586. https://doi.org/10.1145/2858036.2858592

[18] Jason T. Jacques and Per Ola Kristensson. 2019. Crowdworker Economics in the Gig Economy. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–10. https://doi.org/10.1145/3290605.3300621

[19] Farnaz Jahanbakhsh, Justin Cranshaw, Scott Counts, Walter S. Lasecki, and Kori Inkpen. 2020. An Experimental Study of Bias in Platform Worker Ratings: The Role of Performance Quality and Gender. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376860

[20] Phil Jones. 2021. *Work without the worker: labour in the age of platform capitalism*. Verso Books, Brooklyn.

[21] Bernard Koch, Emily Denton, Alex Hanna, and Jacob G. Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. *CoRR* abs/2112.01716. Retrieved from https://arxiv.org/abs/2112.01716

[22] Laura Lascau, Sandy J. J. Gould, Anna L. Cox, Elizaveta Karmannaya, and Duncan P. Brumby. 2019. Monotasking or Multitasking: Designing for Crowdworkers' Preferences. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300649

[23] Eric Loepp and Jarrod T. Kelly. 2020. Distinction without a difference? An assessment of MTurk Worker types. *Research & Politics* 7, 1: 205316801990118. https://doi.org/10.1177/2053168019901185

[24] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44, 1: 1–23. https://doi.org/10.3758/s13428-011-0124-6

[25] Milagros Miceli, Julian Posada, and Tianling Yang. 2021. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *arXiv:2109.08131 [cs]*. Retrieved September 21, 2021 from http://arxiv.org/abs/2109.08131

[26] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 161–172. https://doi.org/10.1145/3442188.3445880

[27] Prayag Narula, Philipp Gutheim, David Rolnitzky, Anand Kulkarni, and Bjoern Hartmann. 2011. MobileWorks: A Mobile Crowdsourcing Platform for Workers at the Bottom of the Pyramid. 4.

[28] Thi Thao Duyen T. Nguyen, Thomas Garncarz, Felicia Ng, Laura A. Dabbish, and Steven P. Dow. 2017. Fruitful Feedback: Positive Affective Language and Source Anonymity Improve Critique Reception and Work Outcomes. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1024–1034. https://doi.org/10.1145/2998181.2998319

[29] Andrew Norton. 2017. *Automation and inequality The changing world of work in the global South*. International Institute for Environment and Development (IIED).

[30] Jessica Pater, Amanda Coupe, Rachel Pfafman, Chanda Phelan, Tammy Toscos, and Maia Jacobs. 2021. Standardizing Reporting of Participant Compensation in HCI: A Systematic Literature Review and Recommendations for the Field. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.3445734

[31] Jorge Ramírez, Burcu Sayin, Marcos Baez, Fabio Casati, Luca Cernuzzi, Boualem Benatallah, and Gianluca Demartini. 2021. On the State of Reporting in Crowdsourcing Experiments and a Checklist to Aid Current Practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2: 1–34. https://doi.org/10.1145/3479531

[32] Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. 2015. We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1621–1630. https://doi.org/10.1145/2702123.2702508

[33] Shruti Sannon and Dan Cosley. 2018. "It was a shady HIT": Navigating Work-Related Privacy Concerns on MTurk. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3170427.3188511

[34] Shruti Sannon and Dan Cosley. 2019. Privacy, Power, and Invisible Labor on Amazon Mechanical Turk. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300512

[35] Saiph Savage, Chun Wei Chiang, Susumu Saito, Carlos Toxtli, and Jeffrey Bigham. 2020. Becoming the Super Turker:Increasing Wages via a Strategy from High Earning Workers. In *Proceedings of The Web Conference 2020*, 1241–1252. https://doi.org/10.1145/3366423.3380200

[36] Morgan Klaus Scheuerman, Emily Denton, and Alex Hanna. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *arXiv:2108.04308 [cs]*. https://doi.org/10.1145/3476058

[37] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2: 1–37. https://doi.org/10.1145/3476058

[38] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1: 1–35. https://doi.org/10.1145/3392866

[39] Ted Shelton. 2013. Business models for the social mobile cloud: transform your business using social media, mobile Internet, and cloud computing. John Wiley & Sons, Hoboken, N.J.

[40] M. Six Silberman and Lilly Irani. 2016. Operating an Online Reputation System: Lessons from Turkopticon: 2008-2015. *Comparative Labor Law & Policy Journal* 37, 1. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2729498

[41] M. Six Silberman, Lilly Irani, and Joel Ross. 2010. Ethics and tactics of professional crowdwork. *XRDS: Crossroads, The ACM Magazine for Students* 17, 2: 39–43. https://doi.org/10.1145/1869086.1869100

[42] Caroline Sinders, Rainbow Unicorn, Cade Diehm, and Ian Ardouin Fumat. 2020. *Technically Responsible Knowledge (TKS)*. Retrieved from https://carolinesinders.com/trk/

[43] Anand Sriraman, Jonathan Bragg, and Anand Kulkarni. 2017. Worker-Owned Cooperative Models for Training Artificial Intelligence. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 311–314. https://doi.org/10.1145/3022198.3026356

[44] Syndicat European Trade Union. ETUC resolution on digitalisation towards fair digital work.

[45] Navid Tavanapour and Eva A. C. Bittner. 2018. The Collaboration of Crowd Workers. In 26th European Conference on Information Systems: Beyond Digitization - Facets of Socio-Technical Change, ECIS 2018, Portsmouth, UK, June 23-28, 2018, 65. Retrieved from https://aisel.aisnet.org/ecis2018_rip/65

[46] Stephen Uzor, Jason T. Jacques, John J Dudley, and Per Ola Kristensson. 2021. Investigating the Accessibility of Crowdwork Tasks on Mechanical Turk. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3411764.3445291

[47] Rama Adithya Varanasi, Aditya Vashistha, Divya Siddarth, Vivek Seshadri, and Kalika Bali. 2022. Feeling Proud, Feeling Embarrassed: Experiences of Low-income Women with CrowdWork. In *In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. https://doi.org/10.1145/3491102.3501834

[48] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2018. BSpeak: An Accessible Voice-based Crowdsourcing Marketplace for Low-Income Blind People. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3173631

[49] Xinyi Wang, Haiyi Zhu, Yangyun Li, Yu Cui, and Joseph Konstan. 2017. A Community Rather Than A Union: Understanding Self-Organization Phenomenon on MTurk and How It Impacts Turkers and Requesters. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2210–2216. https://doi.org/10.1145/3027063.3053150

[50] Mark E. Whiting, Dilrukshi Gamage, Snehalkumar (Neil) S. Gaikwad, Aaron Gilbee, Shirish Goyal, Alipta Ballav, Dinesh Majeti, Nalin Chhibber, Angela Richmond-Fuller, Freddie Vargus, Tejas Seshadri Sarma, Varshine Chandrakanthan, Teogenes Moura, Mohamed Hashim Salih, Gabriel Bayomi Tinoco Kalejaiye, Adam Ginzberg, Catherine A. Mullings, Yoni Dayan, Kristy Milland, Henrique Orefice, Jeff Regino, Sayna Parsi, Kunz Mainali, Vibhor Sehgal, Sekandar Matin, Akshansh Sinha, Rajan Vaish, and Michael S. Bernstein. 2017. Crowd Guilds: Worker-led Reputation and Feedback on Crowdsourcing Platforms. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1902–1913. https://doi.org/10.1145/2998181.2998234

[51] Mark E. Whiting, Grant Hugh, and Michael S. Bernstein. 2019. Fair Work: Crowd Work Minimum Wage with One Line of Code. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1: 197–206.

[52] Huichuan Xia, Yang Wang, Yun Huang, and Anuj Shah. 2017. "Our Privacy Needs to be Protected at All Costs": Crowd Workers' Privacy Experiences on Amazon Mechanical Turk. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW: 1–22. https://doi.org/10.1145/3134748

[53] Rui Yang and Hongbo Sun. 2021. A new simulation framework for crowd collaborations. *International Journal of Crowd Science* 5, 1: 2–16. https://doi.org/10.1108/IJCS-02-2020-0006

[54] 2014. Guidelines for Academic Requesters - Turkopticon. *Turkopticon Blog*. Retrieved from https://blog.turkopticon.net/?page_id=121