# Reputation Agent: Prompting Fair Reviews in Gig Markets

Carlos Toxtli, Angela Richmond-Fuller

HCI Lab, West Virginia University

carlos.toxtli@mail.wvu.edu,angela.richmond@gmail.com

Saiph Savage

HCI Lab, West Virginia University

saiph.savage@mail.wvu.edu

## ABSTRACT

Our study presents a new tool, Reputation Agent, to promote fairer reviews from requesters (employers or customers) on gig markets. Unfair reviews, created when requesters consider factors outside of a worker's control, are known to plague gig workers and can result in lost job opportunities and even termination from the marketplace. Our tool leverages machine learning to implement an intelligent interface that: (1) uses deep learning to automatically detect when an individual has included unfair factors into her review (factors outside the worker's control per the policies of the market); and (2) prompts the individual to reconsider her review if she has incorporated unfair factors. To study the effectiveness of Reputation Agent, we conducted a controlled experiment over different gig markets. Our experiment illustrates that across markets, Reputation Agent, in contrast with traditional approaches, motivates requesters to review gig workers' performance more fairly. We discuss how tools that bring more transparency to employers about the policies of a gig market can help build empathy thus resulting in reasoned discussions around potential injustices towards workers generated by these interfaces. Our vision is that with tools that promote truth and transparency we can bring fairer treatment to gig workers.

## 1 INTRODUCTION

Gig marketplaces are online spaces where almost anyone can contract independent workers (e.g., freelancers) to conduct labor or deliver services in the form of short-term engagements [24]. Gig markets facilitate transactions between strangers, as people typically have to hire and manage a crowd of workers they have never met [10, 98]. Similarly, workers on these platforms often coordinate with other workers[54] and offer services to requesters (customers or employers) who are all unfamiliar to them [83]. These direct interactions between strangers mean that gig markets must have mechanisms through which people can truthfully assess each other, i.e., earning money by entrust their hard labor to strangers [35, 78, 84, 95, 100]. One of the most popular instruments for helping people to assess strangers and to choose who to hire are reputation systems which function by asking individuals to provide feedback on others' work. These systems are generally based on the platform's review metrics. For gig worker's, reputation has become especially important because it is critical for accessing higher-paying jobs [74, 77] or even staying employed [44].

In this context, it is important to understand that in the power dynamics of most gig markets, the platform takes the side of requesters [98] or the platforms manipulate the market to the detriment of the worker [17, 98]. Therefore, if a requester invests time to write a lengthy complaint about a worker (even if the requester is incorrect), the market will side with the requester, potentially leading to unfair termination. For instance, the following advise is from an Uber driver for other gig drivers [27]: *"...[one of the main reasons for Uber to terminate a driver is that the] passenger makes a serious*
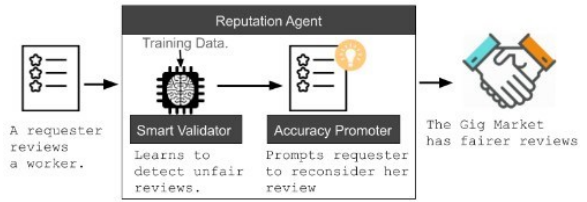


Figure 1: Overview of how Reputation Agent functions.

*complaint about you [the driver]. If a passenger goes out of their way to tell Uber that you were rude, or that you're a bad driver, or that you made them uncomfortable in any way, you can be immediately deactivated without prior notice. You aren't likely to be reactivated after a major passenger complaint..."*

This environment where workers have limited mechanisms to negotiate or even discuss reviews has led workers to distrust gig markets altogether [44, 52]. Thus, it is crucial to ensure that the reviews about workers are fair in order to improve trust and the general operation of gig markets [9]. Fairness within gig markets typically involves ensuring that the policies on the market are transparent, concise, and accessible to workers [42, 96]. However, we argue that fairness is not just about empowering workers to understand the policies of the markets in which they participate; it is also about guiding the requester in the market to evaluate workers based on the market's established policies [33].

Requesters should have a clear understanding of what metrics they should consider and which are inappropriate. Thus, it becomes critical to have mechanisms through which requesters can discriminate between the interactions and labor that workers are expected to control (according to the policies of the market) and those that were outside workers' control [25]. This first type of interaction is known as "mission-critical" and the latter as "non-critical" [103]. Gig marketplaces have historically had difficulties in ensuring that requesters focus on evaluating mission-critical factors [39, 90]. To help, practitioners and researchers have started to investigate different interfaces for facilitating the generation of more mission-critical reviews [37] which use drop-down lists to guide people on what metrics to focus on [54]. However, these interfaces rarely focus on guiding people on the written reviews. As a consequence, unjust written reviews

continue to plague gig markets [13, 72] and have resulted in problematic outcomes, such as termination of workers' accounts. This can eliminate an important source of income [88].

Seeing the need to motivate requesters to write fairer written reviews on gig markets, we present a new novel tool, Reputation Agent, which is an intelligent web plugin that detects when requesters have written reviews that consider factors outside a worker's control. In such cases, Reputation Agent prompts people to reflect and focus on the performance metrics that are actually within the workers' control sphere. We designed Reputation Agent as a web plugin to empower platform maintainers to easily integrate the tool into their existing gig markets without having to change any of their front-end interfaces. Fig 1. presents an overview of how Reputation Agent functions. We conducted a study to evaluate how effective Reputation Agent was in prompting people to generate reviews that focus on metrics within workers' control. In order to investigate our tool in depth, we chose various gig markets (Uber, Upwork, and Grubhub) and recruited 480 reviewers to evaluate gig workers across several scenarios. Across these different gig markets, reviewers working with Reputation Agent were motivated to focus significantly more on metrics that the worker could control instead of metrics outside of the worker's scope.

Our paper contributes a new tool which leverages machine learning for fostering fairer written reviews about gig workers. Our design also provides a novel understanding of requesters who didn't change their reviews for the following reasons: truthfulness, empathy, warning, and agency. Through the ease of implementing our plug-in tool and the understanding gained about why requesters don't change their reviews, we hope that further studies can elaborate on the importance of integrating requesters' decision-making process into their studies to order to achieve fairer reviews for workers. Our discussion: *(i)* focuses on how tools such as Reputation Agent can motivate requesters to write more accurate performance reviews in a manner that provides productive feedback for the crowd market community; and *(ii)* explores how tools, like Reputation Agent, could help to develop empathy and motivate reflections on the type of policies and agencies that people desire within a gig market. Our hope is that systems, like Reputation Agent, can initiate a future environment where workers operate in a fairer, more truthful space; an atmosphere in which all participants have a clearer understanding of the policies and labor conditions of the crowd market. This along with the hope that it will guide a future where gig workers no longer fear unjust termination.

## 2 RELATED WORK

The design of Reputation Agent is based on two main areas: (1) tools for written reviews; and (2) reputation systems.

## 2.1 Gig Marketplaces

Gig markets bring new jobs to the marketplace [16]. However, due to the nontraditional nature of the gig economy, criteria and tools to improve the labor conditions for workers are still necessary for researchers to investigate to ensure a safe and fair working environment for gig workers. [13, 14, 47–49, 53–55, 59, 95, 96]. Gig markets rely heavily on reviews to help requesters identify which workers to hire and help workers ensure fair compensation [61, 92]. It is this reliance on reviews that our study focuses on as bad reviews can pose an obstacle to workers. This is due to the fact that gig markets have been plagued with unfair reviews which contain inaccurate reputation signals about workers. These unfair reviews can ultimately limit workers' future job opportunities and can also result in workers not getting paid or even being terminated from the marketplace. Unfair reviews are generally created because employers have a hard time differentiating the factors within the workers' control and the ones that have little to do with their performance (e.g., when they complain about an Uber driver getting stuck in traffic). However, because market power is typically placed in the hands of employers, a bad worker review can result in the worker losing her entire livelihood. It is important to research how tools can be implemented to protect gig workers [102].

## 2.2 Tools for Written Reviews

Platforms for improving people's written reviews can be divided into two main types: Interface or Artificial Intelligence based.

*2.2.1 Interface Based.* Several interfaces have emerged that focus on driving people to provide better-written reviews about others. One set of these systems has focused on guiding better reviews within educational systems[18, 66]. Cook et al. [22] explored how the use of interfaces that have guiding questions can facilitate the generation of better reviews within project-based learning. In our research, we build on the ideas behind these systems to now imagine interfaces that guide requesters to write fairer reviews about their fellow workers.

*2.2.2 Artificial Intelligence Based.* Another subset of related tools have focused on using artificial intelligence to help reviewers. The work of Krause et al. [64] explored how natural language models could be used to guide designers to provide higher quality reviews about the work of their peers (which was not necessarily fair). Inspired by these ideas, we explored with Reputation Agent how different language models could be used to now guide requesters to write fairer reviews. For Reputation Agent we also used language models to identify when a requester is writing a review that is unfair and guide requesters to write reviews more focused on factors that workers controlled.

Some of the first intelligent tools around reviews were automated methods that inferred the expected reputation scores that people would input based on their written reviews [2, 85]. Others developed sentiment analysis methods to detect the polarity (positive, negative or neutral) of reviews [21, 29]. Sentiment analysis has played an important role in improving the automated analysis and understanding of text reviews [8, 31, 46, 73]. Similarly, developments in deep learning algorithms have further facilitated the automated understanding and even

categorization of marketplace reviews [63, 67]. Deep learning algorithms and other related methods have facilitated automatically detecting more complex metrics aside from sentiment, such as the expected level of helpfulness of a review [97], who was to blame for a car accident based on a car insurgence report [32], health risks in restaurants based on people's reviews on Twitter [89] or detecting biased Amazon reviews [30]. We use inspiration from these intelligent systems to envision how deep learning could be used to automatically flag unfair reviews.

*2.2.3 Fairness In Crowd-Powered Text Reviews.* Within the context of Gig markets, fairness usually relates to the conditions of the workers laboring on these platforms [36]. Graham et al. [43] recently created a framework to score gig markets based on how fair they are to workers. Some of the variables considered were whether the platform paid gig workers the minimum wage and ensured their health and safety at work. Other measures revolved around whether the contracts and policies were transparent, concise, and accessible to workers. Our goal with Reputation Agent, inspired by the latter point, was to facilitate mechanisms through which the policies of a gig market could be presented in a clearer, more conscientious manner. However, our focus was not just on presenting the policies to workers. But rather, facilitating an understanding of policies by requesters, who must judge workers and can, ultimately, have a lasting effect on their future job opportunities. Additionally, we focused on designing a tool that could be easily adapted to any gig market. We believe fairer marketplaces can be constructed by presenting more clearly to employers the roles of workers.

Ensuring fairness in performance evaluations is a common challenge across gig markets [15, 79]. Unfair evaluations can come from individuals or groups [4]. Several systems have implemented different mechanisms to ensure that the evaluations that people generate about others are fair. However, most of these systems operate only at the score or metrics level [91]. These systems, generally, do not take any action to correct "nasty"reviews. But, leaving unfair textual reviews intact can mean that the review can continue to affect the person long after the interaction took place. With Reputation Agent we focus on addressing this problem by detecting unfair reviews, and guiding employers to take action to correct them.

## 2.3    Reputation Systems

A reputation system is any platform that evaluates businesses or peers based on an algorithm or customer rankings, ratings, or written comments [104]. The premise is to have parties rate each other which results in a score. This score should assist other parties in deciding whether or not to continue interacting with that party in the future [58]. To operate effectively, reputation systems require at least three properties: long-lived entities that inspire an expectation of future interactions; capture and distribution of feedback about current interactions (information must be visible in the future); and use of feedback to guide trust decisions [86].

The end goal of reputation systems is to strengthen the quality of markets and communities by providing an incentive for good behavior and quality services, and by sanctioning bad behavior and low-quality services [57]. In order to achieve that goal, some reputation systems have implemented diverse workflows and validations. PowerTrust [105] takes the power-law distribution in user feedback to get a more accurate global reputation. Whitby et al. [101] describe a statistical filtering technique for excluding unfair ratings via a Bayesian reputation system. Notice that prior work focused on improving score based reputations, while our work is based on the foundations of these systems to now assist gig markets in reducing the number of unfair written reputation signals.

*2.3.1 Reputation Systems For Gig Markets.* Within gig markets, reputation systems typically focus on evaluating the different actors involved in the market (workers, requesters and the platform itself)[3]. Reputation systems within the context of gig markets have become a key component for selecting the workers and clients with whom one will collaborate. A worker's income positively correlates with higher reputation scores [40, 52]. Therefore, "bad reviews" can affect worker's access to employment and can overall affect workers' livelihood. Thus, designing accessible tools can promote fairness and enables a shift in the power dynamics [102].

Different interfaces have been introduced to prompt and guide people to write reviews that better match the labor of workers and are potentially more fair and maintain more accurate reputations on the marketplace. For instance, Gaikwad et al. [39] developed Boomerang in Daemo Crowd Market, and explored interfaces that benefited requesters by sharing more accurate information about workers and penalizing requesters who shared inaccuracies. Such mechanisms might not only help workers to obtain better assessments of their work, but it can also help to address the ballot stuffing problem (where people get too many positive reviews, and it thus becomes difficult to assess who is "good"). Our research is inspired by these prior mechanisms to drive fairer reviews and prevent assessments that may unfairly affect the reputation and even the income of gig workers.

## 3    REPUTATION AGENT

We argue that a way to enable fairer reviews in gig markets is via systems that can present requesters with transparent policies that pertain to workers without interrupting their review writing process. This information should only be highlighted in cases when the system identifies that the reviewer has included unfair factors (i.e., factors, per the market's policies, outside a worker's control). For this purpose, our research explores: (1) machine learning techniques that detect when an individual is focusing her review on factors outside the workers' sphere of control; and (2) interfaces that use that information to then prompt the person to reconsider her review in order to refocus on factors within the worker's control. Reputation Agent has two main components: a 'Smart Validator', to detect elements of a review that includes

factors outside a worker's control; and a 'Fairness Promoter', to guide people to focus their review on the factors that were within the worker's control. Figure 1 shows how Reputation Agent enhances existing review forms with these two main parts.

Smart Validator. This component learns to detect when a review has factors outside the worker's scope according to the policies of the market. The Smart Validator has an end-to-end workflow for training a machine learning model. The steps are:

A. PREPARE TRAINING DATA. This piece focuses on collectingreal reviews about gig workers. It functions as a web crawler that collects data from websites, such as SiteJabber and ConsumerAffairs, that share real-world reviews about gig workers. Once the data is collected, the module focuses on labeling each review based on whether it focuses on mission-critical metrics or not (i.e., factors that the worker controlled or not) The labeling is done by analyzing the policies of each gig market.

B. TRAIN AND TEST INTELLIGENT MODEL. Given a set oflabeled reviews (i.e., reviews that are labeled as to whether they are fair or unfair), Reputation Agent uses stratified sampling to split the labeled data into training, test, and validation sets under proportions of 80%, 10%, and 10% (the validation set helps to avoid overfitting). Using Python 3 and the Keras framework with Tensorflow, we trained eight models to learn to recognize reviews that evaluate workers based on mission-critical metrics and non-critical ones. Our goal was to identify the machine learning models which worked the best for different gig markets. For this purpose, we trained different machine learning models which used as feature vectors either word vectors or embeddings:

Word ngram + LR: Logistic regression with word ngrams.

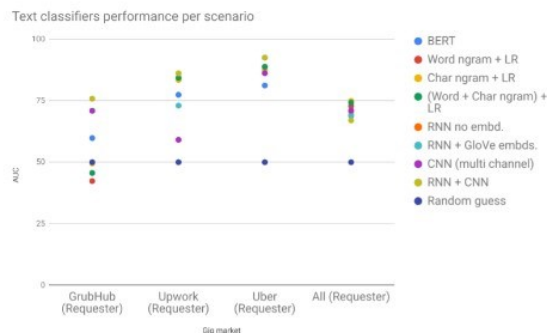Char ngram + LR: Logistic regression with character ngrams.



Figure 2: Text classifier benchmark. Recurrent Neural Networks (RNN) + Convolutional Neural Network (CNN) approach performed better across conditions.

(Word + Char ngram) + LR: Logistic regression with word and character ngrams.

RNN no embedding: Recurrent neural network (bidirectional GRU) without pre-trained embeddings.

RNN + GloVe embedding: Recurrent neural network (bidirectional GRU) with GloVe pre-trained embeddings. CNN (multi-channel): Multi-channel Convolutional Neural Network.

RNN + CNN: Recurrent neural network (Bidirectional GRU) + Convolutional Neural Network.

Google BERT [26]: Bidirectional Encoder Representations from Transformers, is a new method of pre-training language representations which obtains state-of-the-art results on a wide range of Natural Language Processing (NLP) tasks.

We implemented early stopping as a method to stop training once the model performance stops improving on a hold out validation dataset. For the deep learning models, we used a binary crossentropy loss function, ADAM as an optimizer, and a learning rate of 0.001. Fig. 2 presents an overview of the benchmark of the training models (i.e., the figure shows the performance metrics of each model). We note that different machine learning models performed better on different gig markets. However, RNN (Recursive Neural Network, a Deep Learning Algorithm) performed in general the best across all gig markets. This was the reason we eventually choose to utilize this model. After the model has been trained, it is exposed as a REST web service via JSON requests to our front-end interface.
The service is consumed directly by Reputation Agent's Accuracy Promoter and it displays the messages accordingly.

Fairness Promoter. This component displays messages to the reviewer to prompt them to avoid considering factors outside the worker's control based on the policies of the market. It displays the prompt messages that the platform maintainer defines and the messages are triggered based on the predictions from the Smart Validator. The Fairness Promoter is a web plugin for JQuery, a javascript framework, and works as a form validation plugin (commonly used to prevent forms from submitting data that do not fulfill a website's validation or formatting criteria). Reputation Agent's Fairness Promoter is linked to a text control that triggers a request to the Smart Validator every time the text control stops being used by the reviewer. The Fairness Promoter sends the reviewer's current text to the Smart Validator in a JSON format. The Smart Validator analyzes the text and returns the likelihood of whether or not the review is focusing on factors that were within the control of the gig worker. If it is not, the Fairness Promoter then displays its configured messages to prompt reviewers to reconsider their review and focus instead on variables that the worker was able to control.

In our design of the Fairness Promoter we chose for in-form prompting instead of popups. The logic behind this decision is that this design can lead to faster completion times [50]. This is an important decision due to the limited time that customers usually spend evaluating services. Additionally, we considered that some users might have popup blockers in their browsers that could prevent them from seeing the prompt. Therefore, we opted to explore other approaches. Additionally, we chose the prompted message to be shown after the end-user finished writing her review, instead of while she was completing it. We made this decision because prior research has shown that people tend to be

in either a formcompletion-mode or a problem resolution mode [7]. If people are in a form-completion-mode, they tend to ignore alert messages (and hence Reputation Agent would be less effective). Furthermore, we decided to place Reputation Agent's prompting messages close to the review text box since previous work [93] has shown that such placement is more effective than when it is placed on top or at the bottom of the review form. On the other hand, while our prompting messages can be edited by platform maintainers to publish whatever message they desire; we aimed for the initial boxed messages to follow guidelines that prior work has deemed are the most effective. In particular, we follow the design guidelines from Bargas-Avila et al. [6] that stated that promoting messages should be polite, explain the problem, and outline a solution. Our explanation aimed to convey to requesters how their review might be considered unfair based on the policies of the market. We also aimed to briefly explain what type of factors are considered to be outside the control of a worker; and offered people the solution to re-write the parts of their review with those unfair factors.

## 4 EVALUATION

Reputation Agent instantiates our design hypothesis that by flagging reviews with factors outside a worker's control and then presenting to employers what the policies of the gig market highlight as workers' responsibilities, we can prompt fairer assessments of gig workers. To test this hypothesis, we conducted a between-subjects study comparing Reputation Agent with control interfaces. We had participants evaluate a gig worker, given a scenario where the customer had experienced a "bad outcome" on the gig market. However, it was not the fault of the worker being evaluated (i.e., factors outside the worker's control were to blame). We study whether people using Reputation Agent generated fairer reviews than people using control interfaces to review gig workers under the same circumstances. Given that it was also important for us to evaluate our tool within different gig markets (considering that it could be used in diverse niches), we evaluated our tool on marketplaces similar to: Uber, GrubHub and Upwork.

### 4.1 Method

Our study focused on three popular gig markets (Grubhub, Uber, and Upwork). We randomized participants into one of our experimental conditions which represented a particular gig market and interface for reviewing workers (Reputation agent or control). Participants had to imagine they were a gig market customer or employer who had to write a review about a gig worker after experiencing a "bad" outcome on the marketplace (which was not the worker's fault). The scenarios that participants had to consider were:

1) **Uber Scenario.** Participants are passengers in a ride-sharing platform (e.g., Uber) where the driver had followed the recommended GPS route, had a clean car, had picked them up and dropped them off in the correct locations, and was polite. However, due to the heavy traffic, they experienced a delayed trip and had to pay an overpriced fee. The tardiness of their ride resulted in them missing an important meeting with a client and losing a contract.

2) **GrubHub Scenario.** Participants have an important lunch with a client and ordered the meal through an on-demand food delivery platform (GrubHub). The delivery person dropped off her meal on time. However, the meal contained an ingredient that caused the client to have an allergic reaction; thus, making the client very sick (The order had included a request for this ingredient to be removed). Due to the bad experience, the client decided to cancel her contract with our participants.

3) **Upwork Scenario.** Participants used a freelancing platform (Upwork) to hire someone to translate an important report from English to French for a French client. The translation was delivered on time and the translation seemed to be of high quality. However, due to a glitch in the system, the last part of the essay was truncated. Because the customer did not know French they had not realized that the report was truncated. They gave the truncated translation to their French client making a bad impression and losing the contract with the client.

For each of these three gig markets, we trained our tool to detect reviews that involved factors outside a worker's control. For this purpose, for each gig market we: (1) collected 1,000 real-world reviews from SiteJabber (for the three scenarios); (2) had two independent college graduate coders classify each of these reviews into whether they involved worker's performance or factors outside the worker's control. We provided summaries of what factors were considered to be within the worker's control and examples of which ones were not. Coders were also given a link to the policies of each of the three gig markets to better assess the variables that the marketplace considers are under a worker's control. Some explained examples were given to coders to have a common agreement when dealing with ambiguous cases. The two coders agreed on the classification of 94.7% of all the reviews (Cohen's kappa =.86: Strong agreement). We then asked a third college graduate coder to act as a tiebreaker in cases of disagreement. After this step, for all three types of gig markets, we had a labeled set of reviews. The labeled data was provided as input to Reputation Agent's Smart Validator to train its models. Reputation Agent uses stratified sampling to split the labeled data into training, test, and validation sets under proportions of 80%, 10%, and 10% (the validation set helps to avoid overfitting). We implemented early stopping as a method to stop training once the model performance stops improving on a hold

**Upwork**

How likely are you to recommend this freelancer to a friend or a colleague?
⭐⭐⭐⭐⭐⭐⭐⭐⭐⭐

Feedback to freelancer
Availability
⭐⭐⭐⭐⭐

Communication
⭐⭐⭐⭐⭐

Adherence to Schedule
⭐⭐⭐⭐⭐

Skills
⭐⭐⭐⭐⭐

Quality of work
⭐⭐⭐⭐⭐

Cooperation
⭐⭐⭐⭐⭐

Provide a written review about your freelancer
Review to the driver

0 characters

**Control + Rating**

Rating
⭐⭐⭐⭐⭐

What went wrong?
Poor route
Too many pickups
Co-rider behavior
Navigation
Driving
Other

Please provide a written review about your driver.
Review to the driver

0 characters

**GrubHub**

Was the delivery person professional and courteous?
Yes ○   No ○

Was your delivery on time?
Yes ○   No ○

Rate your overall experience
⭐⭐⭐⭐⭐

Provide a written review about your delivery person
Review to the driver

0 characters

Revised Review
Revised Review

0 characters

Submit

**Reputation Agent + Rating**

Our system detected that you likely wrote the review considering things that the driver could not control. Please re-write your review trying to remove these things.
Remember that the traffic, the GPS route, and the rates are not under the driver's control.Please take this into account when reviewing your driver again. You can check out more about Uber's policies for drivers here.

Revised Review
Revised Review

0 characters

Submit

Figure 3: General Interfaces per condition.

out validation dataset. For our deep learning module, we used a binary cross-entropy loss function, ADAM [60] as an optimizer, and a learning rate of 0.001.

Across conditions, participants wrote their review based on their fictional scenario and using one of the four possible interfaces:

1) Control (written text). The end-user is presented with a traditional textbox where they must write a review about the worker.

2) Control + Rating. The end-user is presented with a traditional textbox where they must write a review about the worker, as well as complete traditional 5-star rating questions. These rating questions match the ratings that are currently present in the particular gig market in which the participant is operating (e.g., participants in the Uber scenario were presented with the rating questions that Uber uses to review drivers).
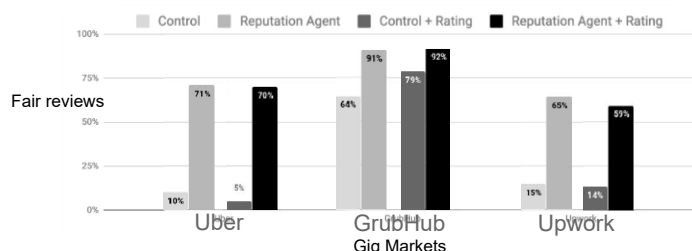
Figure 5: Percentage of fair reviews created per gig market scenario and when using a particular interface.

3) Reputation Agent. The end-user is presented with a traditional textbox where they must write a review about the gig worker while receiving prompting from Reputation Agent.

4) Reputation Agent + Rating. The end-user is presented with a traditional textbox where they must write a review about the gig worker while receiving prompting from Reputation Agent. The end-user is also asked to complete traditional 5-star rating questions about the worker. Here, the rating questions again match the questions present in the given gig market.

Fig. 3 presents a general overview of how the interfaces per each condition looked. Fig. 4 presents the different rating interfaces we considered per gig market. We aimed for these rating interfaces to mimic the ratings that particular gig markets have as we were interested in studying how our tool performed within mainstream settings. Notice that in the Reputation Agent conditions, we stored all the review attempts to analyze how people's behavior changed.

**Rating interfaces**

Uber

Figure 4: Rating interfaces per Gig Market.

Our between-subjects study had 12 conditions that involved three different fictional scenarios (three types of gig markets) where four different interfaces for reviewing workers were used. Each condition had 40 participants. People's participation consisted of writing the review for the worker they were assigned and then completing a follow-up survey to provide feedback about their experiences. Specifically, the survey questioned people about: (1) Who or what was responsible for the bad service they had received on the gig platform? (gig worker, requester, platform algorithms, client, or other) (2) How much fault did each of those actors have? (3) How much did they think that their review would affect the worker's reputation? (4) As a customer of gig platforms, what type of review interface (written or 5-star rating reviews) did they prefer? (5) As a worker or requester of gig markets, what type of reputation mechanism (written or 5-star rating reviews) did they prefer? (6) How much did they feel that the interface helped them to give more accurate feedback about worker's performance? (7) How did their review process (if any) change after completing the review with their interface? Once participants had submitted their review and completed the follow-up survey, we analyzed whether the reviews they submitted were fair, specifically whether they integrated factors outside a gig worker's control or not (to study the effectiveness of Reputation Agent). For this purpose, we had two independent college graduate coders read each of the final reviews that participants generated and categorize whether the review blamed the worker on factors outside the workers' control or not (coders were also given the policies of each gig market to help their categorization, examples and summaries of the policies). The two coders agreed on the classification of 95.1% of all the reviews produced by participants (Cohen's kappa =.87: Strong agreement). In cases where there was disagreement, we asked a third college graduate coder to act as a tiebreaker. In all cases, we categorized the first and last reviews submitted in order to determine how much Reputation Agent lead people to change their reviewing behavior.

We recruited a total of 480 participants using university mailing lists, social media, and via postings on gig markets. Note that these

Fair reviews per interface

are the same methods utilized by prior work to recruit requesters for studies [39]. Important to note is that all our participants had been at least once a requester (employer or customer) on the gig

market to which they were assigned. 55% of our participants were male, and 45% were female. Participants were between the ages of 21 to 70 years old, with the median age being 35. All had at least a High School degree, 59% had a bachelor's degree, 17% a master's degree, and 2% a Ph.D. degree. Some of our participants had been workers at least once on gig platforms: 18% on Uber, 18% on Upwork, and 12% on GrubHub (which is normal given that gig markets allow people to take on both roles.) Participants were paid $2.00 USD to participate, and the study took at most 15 minutes.

## 5  RESULTS

Figure 5 presents across gig markets the number of fair reviews that were generated when using a particular interface. Table 1 presents examples of reviews that were classified as unfair and fair. Across gig markets, the people using Reputation Agent wrote a larger number of fair reviews. In certain scenarios, having people use Reputation Agent leads to an increment of fairer reviews in comparison to when Reputation Agent was not used. For instance, for the Uber scenario when people used the control interface, only 10% of the reviews were fair (i.e., 90% were unfair reviews where people blamed their driver for factors outside her control, e.g., bad traffic.) However, when using Reputation Agent the number of fair reviews increased up to 70%.

We also note that in some scenarios, having Reputation Agent operate with both textbox and numerical rankings lead people to write a slightly higher number of fairer reviews than when using only a textbox (this is the case for the GrubHub scenario). However, we also note that for Upwork and Uber there were a higher number of fair reviews when Reputation Agent operated only with a textbox (and no numerical ranking). To test whether these observed differences were significant or not, we conducted a series of statistical tests. After determining that our data did not meet the normality assumption, we decided to run an omnibus non-parametric Kruskal-Wallis test ($p < .00001$, $H = 81.9303$) and the Mann–Whitney U-tests with Bonferroni correction ($p < .00001$, $z = 9.05126$, $U = 20400$) to identify post hoc effects over conditions. Through this analysis, we found that there was indeed a significant difference in the number of fair reviews that people generated when using Reputation Agent when compared to reviews generated with Table 1: Examples of reviews that were written by study participants and categorized as "unfair" and "fair" across gig market scenarios.

| Unfair | Blame worker for things outside their control |
|---|---|
| Uber | "I am beyond pissed. This driver took a ridiculous route causing a 45min delay and had the nerve to overcharge me on it! I was late for work and obviously, my boss was not pleased. Thanks a lot, Uber." |
| GrubHub | "The delivery person failed to perform one of the basic functions of their job, which was ensuring the product they picked up was correct. He brought lunch to one of my clients and she couldn't even eat it due to allergy concerns. Our food had peanut butter when I clearly stated no peanut butter. Very annoyed. I wasted my time and my money." |
| Upwork | "Working with this person was a pain in the ass! I had a very important proposal to give to an important potential client who only spoke French. As I only speak English I contracted with this worker to translate it for me. They submitted too close to the deadline with little wiggle room, making it impossible to fix or check for any issues. They sent accidentally incomplete work. It cost me a contract due to his negligence. I do not recommend him, as he's not meticulous." |
| Fair | Assessed factors that the worker could control. |

| Uber | "The GPS of my driver leads me to a very congested route today and it took me a lot more time and money to get to work. But these things happen. I could have made the same choice driving myself so the driver can't be blamed for something out of their control. Driver and his car were very nice. I am however very unhappy with the service and will be immediately unsubscribing soon." |
|---|---|
| GrubHub | "The delivery person got the food to me on time. It was not their fault that the order was wrong. The order was wrong because of the people at the restaurant. They messed up what was suppose to be a great lunch. The driver did all he was supposed to do and I give him a good review for that.." |
| Upwork | "I thought the worker did a good job and the work was presented well, it is a shame however that the system failed at the last minute and I got an incomplete submission." |

Table 2: Corrected reviews after Reputation Agent prompted participants. 2 out of every 3 reviews were corrected.

| Gig market | Condition | Corrected reviews |
|---|---|---|
| Uber | Reputation Agent | 66.6% |
| | Reputation Agent + Rating | 70.2% |
| GrubHub | Reputation Agent | 80.9% |
| | Reputation Agent + Rating | 73.3% |
| Upwork | Reputation Agent | 57.8% |
| | Reputation Agent + Rating | 53.4% |

traditional interfaces. In other words, we found that participants are significantly less likely to write unfair reviews when using Reputation Agent than with normal interfaces.

We also investigated how much Reputation Agent helped people to start changing their reviewing behavior. For this purpose, we measured the number of people who, while using Reputation Agent, changed their original review from being unfair to fair (see Table 2). On average, across gig markets, Reputation Agent was able to convert two out of every three of the reviews that were originally unfair to fair (67.1%). For when people decided to not change their unfair review, we analyzed the reasons for this behavior by observing what they stated in the survey. This analysis can help us to identify some of the challenges that Reputation Agent has in ensuring fairness in gig markets. We used open coding to extract initial concepts from people's responses [82]. We aimed for these initial concepts to consider the themes that related work had derived on people's motivations for writing certain types of reviews [41, 68]. Next, we discussed these initial concepts as a group to iterate on them and created a codebook (list of themes).

The codebook with examples was shared with two coders that agree in the 91% of the reasons (Cohen's kappa =.81: Strong agreement) and a third college graduate coder to act as a tiebreaker in cases of disagreement. We detected the following categories describing participants' reasons for not changing their reviews:

5.0.1 Truthfulness. Some reviewers (21% of all reviewers who chose not to change their review after prompts from Reputation Agent) felt that changing their review implied lying about how they felt and wanted to keep their review as it was because it was truthful. Examples of their reasoning:

"I always give honest and detailed reviews and will continue to do so. This interface will not impact on my reviewing process. I will always stand by how I have done things in the past: with the truth!" Uber Reviewer 44.

"Why on earth would my process change? [...] if I get bad work, I'm going to leave an honest review regardless of whether it was the "platform" or the worker's fault...", Grubhub Reviewer 15.

"...I like being honest and factual with reviews. That's why I don't like changing my reviews..." Upwork Review 43.

5.0.2 Agency. These reviewers (25%) felt that although a gig market's policies might dictate that certain actors were not to blame, they believed that such actors should have had more agency in their decisions despite the policies of the gig market.

"Drivers should be able to tell which routes are clean by instinct based on the day and the time of the day without even looking at the GPS. The drivers should know the city that they are driving very well..." Uber Reviewer 23.

"Well the delivery guy didn't listen to a word I said so now my client can't eat the meal. If he has any type of peanut in their food they can go into anaphylactic shock and die. That is not what I want for their lunch? Is that what you want them to have for lunch? Death? The delivery guy really needs to be responsible for this." Grubhub Reviewer 32.

"...Let's be honest, the worker did not perform as well as expected. She missed that some words were cut off. The bottom line is that she needs to learn to handle herself responsibly in the world." Upwork Reviewer 19.

5.0.3 Empathy. These reviewers (36%) felt they needed more time to analyze the scenario before changing their review. They appeared to have empathy for all actors involved and wanted to truly understand their situation before changing their review.

"I like to think about all the circumstances before writing reviews. I like to use empathy. In my future review I will probably be a bit less harsh on drivers. I will think about the driver and how they treated me as well. But before I make those change I will try to calm down first..." Uber Reviewer 21.

"...I wanted to focus on the bad aspects of my meal, but then I realized I was supposed to focus on the deliverer only. So I switched it and focused on the delivery worker instead. I try to take all factors I am aware of into account when writing my review, and wait a bit so that any emotions associated with the work would not affect my review. I think I would wait to re-write my review so I am not angry and really think about the worker..." Grubhub Reviewer 5.

"I like to take software and platform issues into account before completely blasting a worker in a review (either with stars or a written review). I won't change my review now because I have to stop and think about things. Instead of just getting mad and going off on the worker immediately. I would also want to have more communication between worker and purchaser before the review process, and have a way to discuss the review if either party truly found it to be in error." Upwork Reviewer 39.

5.0.4 Warning. Some participants (13%) maintained their review because it was important for them to have a space where they could caution others about what they experienced. They did not care about whether they blamed the incorrect person.

*"...I just detailed the problems I encountered with the driver. I wanted others to know about my issue so that it doesn't happen to them as well."* Uber Reviewer 49.

*"...I usually never provide reviews for delivery people [...] But in this case the service was exceptionally bad and my experience would serve as a cautionary tale to others. So that is why I can't change it [the review]"* GrubHub Reviewer 5.

*"...I think my review would let people know of the risks about using this worker/platform. They could potentially avoid situations like the one that I was in. It's important for me to keep my review to warn others...."* Upwork Reviewer 20.

67% of the people using Reputation Agent reported that they felt that the interface helped them to be more aware of inaccuracies in their reviews. Participants across conditions reported that they felt their review would affect worker's reputation. People in the control condition had the perception that their review would be the most harmful (mean 3.9 of 5). This is notable when compared to people using Reputation Agent who on average thought that their review would not be as harmful (mean 2.9 of 5).

## 6 DISCUSSION

Our experiments demonstrated the potential of using intelligent web plugins to detect unfair reviews on gig markets, and then prompt fairer assessments by presenting micro-information about gig workers' conditions and policies. Across different marketplaces, the majority of people using Reputation Agent ended up writing fairer reviews. Our study provides a novel insight into how marketplaces could use this type of smart web plugin to bring more fairness to workers. In this section, we discuss opportunities and challenges we see with Reputation Agent, and highlight design implications for future systems that operate within the gig marketplace.

### 6.1 Building Empathy In Gig Markets

Taking empathy into account in the human-centered design process can align designers with the values and needs of people who may use the platforms [12]. Mencl et al. defines empathy as "a positive moral emotion that aids reasoning[80]." Our study highlighted that prompting people to reconsider their reviews and reason more deeply about the worker and what her actual job was, helped reviewers to be fairer.

While all our participants considered that their reviews would have an impact on workers' lives, the level of harm that people attributed to their review varied across conditions. People using the control (text only) interface tended to believe their reviews were the most harmful while people using the Reputation Agent + Rating interface felt they were doing the least harm to workers. The "tension between reason and emotion when making decisions[38]" allows us to see the benefit of a tool such as Reputation Agent in prompting requesters to reconsider their written review. Thus, our results highlight that providing more

metrics and guidance helped people feel as if they were doing less harm to workers while still submitting a review that was accurate.

We see tools like Reputation Agent as a way to help requesters have a more humane perspective of workers by providing more transparency and awareness of what the current labor conditions are in gig markets. Through Reputation Agent we offer a way in which requesters can be guided to better understand the actual job expectations for workers. We believe that through this transparency and highlighting of boundaries denoting gig workers' labor that we will be able to build more consideration for workers within gig markets [55]. Several of our participants who Reputation Agent prompted to change their reviews discussed how the tool helped them to better understand workers' conditions.

From our study, we also identified that there were cases where people even after being prompted by Reputation Agent, refused to change their review at all. Many of these individuals were people who felt that workers needed to have more agency. For instance, in the Uber case, some passengers believed that their Uber driver should not have followed the recommended GPS route, but instead selected a shortcut and better route. These individuals blamed their Uber drivers for not taking the initiative and knowing enough of their city to understand that the GPS algorithm was wrong. We believe that in these cases, it might be worth designing interventions where the policies and responsibilities of gig workers are explained in detail to these individuals from the outset. We believe there is an opportunity in using systems like Reputation Agent as a way to create more empathy between requesters and workers. Additionally, it might be worth explaining to workers the perspectives of these requesters in order to facilitate their understanding of why certain requesters might expect them to not always follow an algorithm and be more "proactive." In these cases, we visualize platforms that do not penalize workers for not being proactive (i.e., by following the instructions of the algorithm), but rather help workers open their minds to other perspectives and help them to see that having more agency in their decision process could provide growth opportunities, e.g., to eventually become a manager.

### 6.2 Supporting Reflection In Gig Markets

Our study contained 46 individuals who refused to modify their review even after being prompted. We considered it important to understand the reasons these individuals had for not changing their written assessments. In some cases, requesters used the review process as a chance to communicate to the platform that it would be more efficient (in terms of time and cost) if the worker was allowed more agency. We believe there are benefits to gig markets when they understand the type of agency that requesters want to see in the market. Platforms could consider alternate methods/systems for capturing this type of feedback in order to protect workers.

Williams et al. found that tools that are based on only distributing ratings and reviews for task choosing decisions usually tend to create fragmentation and discrimination affecting

the platform's fairness [102]. We argue that it is important to tie tools like Reputation Agent with platforms focused on driving citizen discussions and citizen reflection [75]. On this point, it is important to consider the findings of Li et al. concerning the influence that embodied conversational agents (ECAs) have on persuading people to consider feedback that is offered them [69]. "The use of agents that resemble users" might be the necessary factor that allows requesters to consider the promptings of the Reputation Agent to be more valid. We must also ask: what type of agencies should we expect from the different actors of the market and why is it important that we expect such agency from them? Further research is needed to investigate the type of interfaces and workflows that could be used to incentivize and guide quality reflections about what people expect from workers, requesters, and the different policies of a marketplace. This type of system could uncover pain points that exist in current crowd markets and where policy changes might be needed. Our study also highlighted another reason why requesters did not desire to change their review: the importance for them to use the space to truthfully share their experiences. In the widely cited paper "The Market for Goods and the Market for Ideas" [20], it is argued that in the market for goods (i.e., the market where consumer goods and services are exchanged), government regulation is desirable; whereas in the market for ideas (i.e., the market where opinions or beliefs are interchanged), government regulation should be limited. Online reviews can be seen as something that delivers both "goods" and "ideas". On one hand, having a person write an objective review of the work someone did could be seen as if they are delivering a good. The good, in this case, corresponds to the overall assessment of the labor that the worker did. This assessment not only helps the market better contextualize and measure the labor that is being produced [56], it can also boost the SEO of the marketplace [94]. Thus, helping it appear higher in the results of search engines and ultimately bring in more customers [87]. The review can also help the worker get better credentials, access higher pay, and more requesters (i.e., the review might persuade other requesters to hire the worker). Reviews as goods deliver services to the marketplace, workers, and even other requesters. However, reviews also have the capacity to deliver opinions and beliefs, and hence can also belong within the market of ideas. Thus, we see value in being able to actively regulate the activities that belong to the market for goods, while permitting freedom of expression for activities that relate to the market for ideas [20]. Reputation Agent offers an advancement towards this area by providing a way to regulate reviews within the market for goods and flagging reviews that might pertain more within the markets of ideas. Future work could pursue this avenue to design review interfaces to express both forms of reviews.

## 6.3 Rating Systems and Fairness

The importance of rating systems and fairness is an essential element in gig markets, whether it pertains to rating workers or rating requesters[28, 54, 95]. Creating a fair working environment with structures designed to protect workers' rights to receive fair

compensation for their labor ensures the reputation and success of gig markets[1, 11, 35, 81, 95] Thus, devising a tool for gig markets to implement in order to ensure fair reviews for workers brings us one step closer to achieving this.

To this purpose, our findings that people tended to write fairer reviews with Reputation Agent when working with the written interface is an important addition to the tools available to gig markets. In our study, we also discovered that people tended to write a larger number of unfair reviews when Reputation Agent was tied with numerical ratings. This was specifically the case with Uber
and Upwork, where having numerical ratings tied with a written review, led to a larger number of unfair reviews than when working with just Reputation Agent and a written interface. Upon closer inspection, we identified that the problem was the fact that the rating systems of these gig markets did not distinguish worker's performance from factors outside the worker's control. For instance, when assessing a driver's rating on Uber, the market provides a
list of possible issues and presents "poor route" as an option even though Uber policies outline that drivers should always follow the recommend GPS route (unless explicitly instructed otherwise by the passenger). As a result, several participants selected "poor route" as an issue and then wrote lengthy reviews blaming the driver for the traffic (despite the prompts from Reputation Agent). Similarly, we noted that markets which differentiated between metrics pertaining to workers vs the platform, led people to generate fairer reviews. For instance, GrubHub has a rating system that differentiates between these two types of metrics, and we saw there was a decrease in the number of unfair reviews generated by participants. We see then the necessity of gig markets to not only incorporate tools that promote fairer reviews, but they themselves must also clarify and communicate the metrics that pertain to the workers.

Unfair reviews may also be the product of biases which do not necessarily reflect a worker's performance, i.e., when a worker gets more positive ratings than expected given the service she provided. These types of reviews can be influenced by cognitive biases such as confirmation bias [62], driven by having prior beliefs; anchoring effect [19], relying more on the first piece of information offered and hence the current performance does not matter; or perception bias [45], motivated by how others might perceive you as the reviewer. Reputation Agent provides the opportunity to educate people about possible biases they might have and how those might be impacting their reviews. Here, we envisage that Reputation Agent's prompts might provide information about biases in addition to information about the policies of gig markets. Future research could focus on personalized feedback according to personality or cultural biases that might exist [51, 65, 70, 76, 106].

## 6.4 Feasibility And Maintainability

Through our controlled experiments, we identified that
Reputation Agent was able to lead requesters to generate fairer

reviews than when they worked with the control interfaces. However, to accomplish these results, it was necessary to have labeled data for each gig market on what constitutes fair reviews and what constitutes unfair reviews. While the labeled data sample that we used was relatively small in comparison to the large number of reviews that are generated on these marketplaces daily [5], it is possible that new gig markets might have a difficult time collecting and labeling review data for Reputation Agent.

We have released our system[1] to help gig markets easily adopt and use our tool. Additionally given that Reputation Agent can be easily implemented as a validation module, Platform maintainers could change their front-end review interface without having to worry about Reputation Agent suddenly not working. Reputation Agent's deep learning nature makes it so that if a gig market changes its policies, platform maintainers with minimum knowledge in artificial intelligence can easily re-train Reputation Agent to be updated with the changes [34, 99]. In our website we have shared training examples for Reputation Agent's learning module so that website maintainers can easily start using our tool.

## 6.5 Key Design Considerations

Our investigation unraveled design considerations for technology to support the generation of fairer interactions on gig markets.

*6.5.1* Tools for Learning about Gig Market Policies. Reputation System can be seen as a tool that helps highlight the policies of a gig market. For instance, when people are writing a review for Uber, Reputation Agent shares Uber's policies. Future work could explore how heuristics and hard-coded rules can lead people to better understand the policies of a marketplace and comprehend what falls under worker vs. platform responsibilities. The visualization of different privacy policy representations can improve the understanding of the different actors [71].

*6.5.2* Tools for Better Moderation. Integrating artificial intelligence (AI) into gig markets can go beyond flagging unfair reviews [23]. AI can also be intermixed with human moderators to facilitate a better understanding of the perspectives of requesters. For instance, given that Reputation Agent is able to store all the review attempts that people make, the system could detect cases when even after prompting the end-user to reconsider her review she still kept everything the same. In which case, the system could trigger an alert to human moderators to take a closer look at the review. We view Reputation Agent as tool that can alleviate moderators' labor. The sustainability and self-management of the tool also depend on the neutrality of the training data. Human-in-the-loop mechanisms can allow different actors to have agency giving everyone decision power, not just the few who can code [102] to define how the tool is learning and taking the decisions.

## 7 LIMITATIONS AND FUTURE WORK

The insights from this work are limited by the methodology and population we studied. Our controlled experiment allowed us to begin understanding how users engage with Reputation Agent. Although, we cannot extrapolate on how people would respond if this approach were implemented in a field deployment with conditions such as limited time and reduced willingness to reconsider their reviews. Our attempt to counter this issue was by implementing interfaces and creating scenarios that mimicked various gig markets and circumstances. However, future work could benefit from analyzing how systems like Reputation Agent are used when people are on the go and suffer from time constraints. While the scenarios we studied resembled very specific real-world situations, our results might not yet generalize to populations at large or to different types of situations. Further analysis is needed to understand how studies that leverage real gig market actors and Reputation Agent play out in helping users to give more objective reviews.

Reputation Agent was designed to limit the amount of extra interface controls that platform maintainers would have to implement. The aim of this work is to provide a smart validation mechanism for existing interfaces, i.e., easy to implement and not invasive. Future work could also explore how adaptations in the workflow and interface controls, such as a separated textbox for the reviews that are generated with Reputation Agent, could lead to reducing unfair reviews. This work explored the effect of using Reputation Agent in two settings: with ratings tied to text reviews and with just written text reviews (without any ratings). We studied whether in these settings Reputation Agent could guide people to change their reviews to fair ones. We choose to focus on written reviews in which having one bad written review could not only lead to a worker having her reputation jeopardized, but also having her account terminated. Future work could explore how integrating fairness validators might also influence the numerical ratings that people give to workers. Our work replicates the current conditions of gig markets, where people are never initially prompted or reminded to be fair in their reviews, i.e., Reputation Agent prompts only when unfair reviews are given. We established this setting because we considered that customers would likely be busy individuals who simply wanted their service delivered. Therefore, constant reminders of the gig market's policies could be considered invasive. If they have not written an unfair review, it might not need highlighting. Future work could explore how promoting fairness throughout different points in time (e.g., directly when starting to write the review or at the end) can lead to fairer reviews. Our study may also have novelty effects that need to be studied through longitudinal studies. Future work could explore how longitudinal studies can promote fair reviews over time. This was a controlled experiment and not a deployment, i.e., there was never money at stake and no real harm done to the worker. Future work can compare how our results differ from deployments in the real world.

---

[1] https://research.hcilab.ml/reputationagent

## REFERENCES

[1] Duncan Adam, Mitchell Bremermann, Jessica Duran, Francesca Fontanarosa, Birgit Kraemer, Hedvig Westphal, Annamaria Kunert, and L TÖNNES LÖNNROOS. 2016. Digitalisation and working life: lessons from the Uber cases around Europe. *EurWORK| European Observatory of Working Life* 25 (2016).

[2] Georgios Alexandridis, Thanos Tagaris, Giorgos Siolas, and Andreas Stafylopatis.
2019. From Free-text User Reviews to Product Recommendation Using Paragraph Vectors and Matrix Factorization. In *Companion Proceedings of The 2019 World Wide Web Conference (WWW '19)*. ACM, New York, NY, USA, 335–343. https:
//doi.org/10.1145/3308560.3316601

[3] Mohammad Allahbakhsh, Aleksandar Ignjatovic, Boualem Benatallah, Elisa Bertino, Norman Foo, et al. 2012. Reputation management in crowdsourcing systems. In *8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*. IEEE, 664–671.

[4] Mohammad Allahbakhsh, Aleksandar Ignjatovic, Boualem Benatallah, Norman Foo, Elisa Bertino, et al. 2014. Representation and querying of unfair evaluations in social rating systems. *Computers & Security* 41 (2014), 68–88.

[5] Anna Baj-Rogowska. 2017. Sentiment analysis of Facebook posts: The Uber case. In *2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS)*. IEEE, 391–395.

[6] Javier A Bargas-Avila, O Brenzikofer, SP Roth, AN Tuch, S Orsini, and K Opwis. 2010. Simple but crucial user interfaces in the world wide web: introducing 20 guidelines for usable web form design. In *User interfaces*. IntechOpen.

[7] Javier A Bargas-Avila, Glenn Oberholzer, Peter Schmutz, Marco de Vito, and Klaus Opwis. 2007. Usable error message presentation in the World Wide Web: Do not show errors right away. *Interacting with Computers* 19, 3 (2007), 330–341.

[8] Alberto Bartoli, Andrea De Lorenzo, Eric Medvet, Dennis Morello, and Fabiano Tarlao. 2016. " Best Dinner Ever!!!": Automatic Generation of Restaurant Reviews with LSTM-RNN. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 721–724.

[9] Andy Beal and Judy Strauss. 2009. *Radically transparent: Monitoring and managing reputations online*. John Wiley & Sons.

[10] Benjamin B Bederson and Alexander J Quinn. 2011. Web workers unite! addressing challenges of online laborers. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. ACM, 97–106.

[11] Christiane Benner. 2014. Amazonisierung oder Humanisierung der Arbeit durch Crowdsourcing. *Crowdwork–zurück in die Zukunft* (2014), 289–300.

[12] Cynthia L. Bennett and Daniela K. Rosner. 2019. The Promise of Empathy: Design, Disability, and Knowing the "Other". In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article Paper 298, 13 pages. https:
//doi.org/10.1145/3290605.3300528

[13] Alan Benson, Aaron Sojourner, and Akhmed Umyarov. 2019. Can Reputation Discipline the Gig Economy?: Experimental Evidence From an Online Labor Market. *Benson, Alan, Aaron Sojourner, and Akhmed Umyarov* (2019).

[14] Janine Berg, Marianne Furrer, Ellie Harmon, U Rani, and M Six Silberman. 2018. *Digital labour platforms and the future of work: Towards decent work in the online world*. International Labour Office Geneva.

[15] Ria Mae Borromeo, Thomas Laurent, Motomichi Toyama, and Sihem AmerYahia. 2017. Fairness and Transparency in Crowdsourcing.. In *EDBT*. 466–469.

[16] Andrea Broughton, Rosie Gloster, Rose Marvell, Martha Green, Jamal Langley, and Alex Martin. 2018. The experiences of individuals in the gig economy. (2018).

[17] Ryan Calo and Alex Rosenblat. 2017. The taking economy: Uber, information, and power. *Colum. L. Rev.* 117 (2017), 1623.

[18] Julia Cambre, Scott Klemmer, and Chinmay Kulkarni. 2018. Juxtapeer: Comparative peer review yields higher quality feedback and promotes deeper reflection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 294.

[19] Andrea Caputo. 2013. A literature review of cognitive biases in negotiation processes. *International Journal of Conflict Management* 24, 4 (2013), 374–398.

[20] Ronald Harry Coase. 1974. The market for goods and the market for ideas. *The American Economic Review* 64, 2 (1974), 384–391.

[21] Anaïs Collomb, Crina Costea, Damien Joyeux, Omar Hasan, and Lionel Brunie. 2014. A study and comparison of sentiment analysis methods for reputation evaluation. *Rapport de recherche RR-LIRIS-2014-002* (2014).

[22] Amy Cook, Jessica Hammer, Salma Elsayed-Ali, and Steven Dow. 2019. How Guiding Questions Facilitate Feedback Exchange in Project-Based Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 138.

[23] Peng Dai, Daniel Sabby Weld, et al. 2011. Artificial intelligence for artificial artificial intelligence. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

[24] Valerio De Stefano. 2015. The rise of the just-in-time workforce: On-demand work, crowdwork, and labor protection in the gig-economy. *Comp. Lab. L. & Pol'y J.* 37 (2015), 471.

[25] Chrysanthos Dellarocas. 2005. Reputation mechanism design in online trading environments with pure moral hazard. *Information systems research* 16, 2 (2005), 209–230.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[27] H. Dough. [n.d.]. Fired From Uber: Why Drivers Get Deactivated, & How to Get Reactivated - Ridesharing Driver. https://www.ridesharingdriver.com/fireduber-drivers-get-deactivated-and-reactivated/. (Accessed on 10/11/2019).

[28] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. ACM, 1013–1022.

[29] Hui Du, Xueke Xu, Xueqi Cheng, Dayong Wu, Yue Liu, and Zhihua Yu. 2016. Aspect-Specific Sentimental Word Embedding for Sentiment Analysis of Online Reviews. In *Proceedings of the 25th International Conference Companion on World Wide Web (WWW '16 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 29–30. https://doi.org/10.1145/2872518.2889403

[30] Elshrif Ibrahim Elmurngi and Abdelouahed Gherbi. 2018. Unfair Reviews Detection on Amazon Reviews using Sentiment Analysis with Supervised Learning Techniques. *JCS* 14, 5 (2018), 714–726.

[31] Ahmad Khamis Elshenawy, Steele Carter, and Daniela Braga. 2016. It's Not Just What You Say, But How You Say It: Muiltimodal Sentiment Analysis Via Crowdsourcing. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

[32] Dominique Estival and Françoise Gayral. 1995. An NLP Approach to a Specific Type of Texts: Car Accident reports. *arXiv preprint cmp-lg/9502032* (1995).

[33] Ming Fan, Yong Tan, and Andrew B Whinston. 2005. Evaluation and design of online cooperative feedback mechanisms for reputation management. *IEEE Transactions on Knowledge and Data Engineering* 17, 2 (2005), 244–254.

[34] Armando Fandango. 2018. *Mastering TensorFlow 1. x: Advanced machine learning and deep learning concepts using TensorFlow 1. x and Keras*. Packt Publishing
Ltd.

[35] Alek Felstiner. 2011. Working the crowd: employment and labor law in the crowdsourcing industry. *Berkeley J. Emp. & Lab. L.* 32 (2011), 143.

[36] Christian Fieseler, Eliane Bucher, and Christian Pieter Hoffmann. 2019. Unfairness by design? The perceived fairness of digital labor on crowdworking platforms. *Journal of Business Ethics* 156, 4 (2019), 987–1005.

[37] Apostolos Filippas, John J Horton, and Joseph Golden. 2017. Reputation in the long-run. (2017).

[38] Chris D Frith and Tania Singer. 2008. The role of social cognition in decision making. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363, 1511 (2008), 3875–3886.

[39] Snehal Gaikwad, Durim Morina, Rohit Nistala, Megha Agarwal, Alison Cossette, Radhika Bhanu, Saiph Savage, Vishwajeet Narwal, Karan Rajpal, Jeff Regino, et al. 2015. Daemo: A self-governed crowdsourcing marketplace. In *Adjunct proceedings of the 28th annual ACM symposium on user interface software & technology*. 101–102.

[40] Alessandro Gandini, Ivana Pais, and Davide Beraldo. 2016. Reputation and trust on online labour markets: The reputation economy of elance. *Work Organisation, Labour and Globalisation* 10, 1 (2016), 27–43.

[41]  Joseph K Goodman and Gabriele Paolacci. 2017. Crowdsourcing consumer research. *Journal of Consumer Research* 44, 1 (2017), 196–210.

[42]  Mark Graham and Jamie Woodcock. 2018. Towards a fairer platform economy: introducing the Fairwork Foundation. *Alternate Routes* 29 (2018).

[43]  Mark Graham, Jamie Woodcock, Richard Heeks, Sandra Fredman, Darcy Du Toit, Jean-Paul van Belle, Paul Mungai, and Abigal Osiki. 2019. The Fairwork Foundation: Strategies for Improving Platform Work. In *Weizenbaum Conference*. DEU,
    8.

[44]  Mary Gray and Siddharth Suri. 2019. Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.

[45]  Jerald Greenberg. 1991. Motivation to inflate performance ratings: Perceptual bias or response bias? *Motivation and Emotion* 15, 1 (1991), 81–97.

[46]  Emitza Guzman and Walid Maalej. 2014. How do users like this feature? a fine grained sentiment analysis of app reviews. In *2014 IEEE 22nd international requirements engineering conference (RE)*. IEEE, 153–162.

[47]  Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 449.

[48]  Ellie Harmon and M Six Silberman. 2018. Rating working conditions on digital labor platforms. *Computer Supported Cooperative Work (CSCW)* 27, 3-6 (2018), 1275–1324.

[49]  Paul Hitlin. 2016. Research in the crowdsourcing age, a case study. *Pew Research Center* 11 (2016).

[50]  Knut Åge Hofseth, Lars Kristian Haga, Vegard Sørlie, and Frode Eika Sandnes. 2019. Form Feedback on the Web: A Comparison of Popup Alerts and InForm Error Messages. In *Innovation in Medicine and Healthcare Systems, and Multimedia*. Springer, 369–379.

[51]  Eileen A Hogan. 1987. Effects of prior expectations on performance ratings: A longitudinal study. *Academy of Management Journal* 30, 2 (1987), 354–368.

[52]  John Horton and Joseph Golden. 2015. Reputation inflation: Evidence from an online labor market. *Work. Pap., NYU* 1 (2015).

[53]  John J Horton. 2011. The condition of the Turking class: Are online employers fair and honest? *Economics Letters* 111, 1 (2011), 10–12.

[54]  Lilly C Irani and M Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 611–620.

[55]  Lilly C Irani and M 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*. ACM, 4573–4586.

[56]  Srikanth Jagabathula, Lakshminarayanan Subramanian, and Ashwin Venkataraman. 2014. Reputation-based worker filtering in crowdsourcing. In *Advances in Neural Information Processing Systems*. 2492–2500.

[57]  Audun Jøsang and Jennifer Golbeck. 2009. Challenges for robust trust and reputation systems. In *Proceedings of the 5th International Workshop on Security and Trust Management (SMT 2009), Saint Malo, France*. Citeseer, 52.

[58]  Audun Jøsang, Roslan Ismail, and Colin Boyd. 2007. A survey of trust and reputation systems for online service provision. *Decision support systems* 43, 2 (2007), 618–644.

[59]  Toni Kaplan, Susumu Saito, Kotaro Hara, and Jeffrey P Bigham. 2018. Striving to Earn More: A Survey of Work Strategies and Tool Use Among Crowd Workers.. In *HCOMP*. 70–78.

[60]  Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[61]  Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1301–1318.

[62]  Joshua Klayman. 1995. Varieties of confirmation bias. In *Psychology of learning and motivation*. Vol. 32. Elsevier, 385–418.

[63]  Marios Kokkodis. 2012. Learning from Positive and Unlabeled Amazon Reviews: Towards Identifying Trustworthy Reviewers. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12 Companion)*. ACM, New York, NY, USA, 545–546. https://doi.org/10.1145/2187980.2188119

[64]  Markus Krause, Tom Garncarz, JiaoJiao Song, Elizabeth M Gerber, Brian P Bailey, and Steven P Dow. 2017. Critique style guide: Improving crowdsourced design feedback with a natural language model. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 4627–4639.

[65]  Frank Krzystofiak, Robert L Cardy, and Jerry Newman. 1988. Implicit personality and performance appraisal: The influence of trait inferences on evaluations of behavior. *Journal of Applied Psychology* 73, 3 (1988), 515.

[66]  Chinmay E Kulkarni, Michael S Bernstein, and Scott R Klemmer. 2015. PeerStudio: rapid peer feedback emphasizes revision and improves performance.
    In *Proceedings of the second (2015) ACM conference on learning@ scale*. ACM, 75–84.

[67]  Srijan Kumar, Bryan Hooi, Disha Makhija, Mohit Kumar, Christos Faloutsos, and VS Subrahamanian. 2017. FairJudge: Trustworthy user prediction in rating platforms. *arXiv preprint arXiv:1703.10545* (2017).

[68]  Steven D Levitt and John A List. 2007. What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic perspectives* 21, 2 (2007), 153–174.

[69]  Ian Li, Jodi Forlizzi, Anind Dey, and Sara Kiesler. 2007. My agent as myself or another: effects on credibility and listening to advice. In *Proceedings of the 2007 conference on Designing pleasurable products and interfaces*. 194–208.

[70]  Ji Li and Leonard Karakowsky. 2001. Do we see eye-to-eye? Implications of cultural differences for cross-cultural management research and practice. *The Journal of Psychology* 135, 5 (2001), 501–517.

[71]  Heather Richter Lipford, Jason Watson, Michael Whitney, Katherine Froiland, and Robert W. Reeder. 2010. Visual vs. Compact: A Comparison of Privacy Policy Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in
    Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1111–1114. https://doi.org/10.1145/1753326.1753492

[72]  Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. 2010. Exploiting Social Context for Review Quality Prediction. In *Proceedings of the 19th
    International Conference on World Wide Web (WWW '10)*. ACM, New York, NY, USA, 691–700. https://doi.org/10.1145/1772690.1772761

[73]  Washington Luiz, Felipe Viegas, Rafael Alencar, Fernando Mourão, Thiago Salles, Dárlinton Carvalho, Marcos Andre Gonçalves, and Leonardo Rocha. 2018. A Feature-Oriented Sentiment Rating for Mobile App Reviews. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1909–1918. https://doi.org/10.1145/3178876.3186168

[74]  Ning F Ma, Chien Wen Yuan, Moojan Ghafurian, and Benjamin V Hanrahan. 2018. Using Stakeholder Theory to Examine Drivers' Stake in Uber. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM,
    83.

[75]  Narges Mahyar, Michael R James, Michelle M Ng, Reginald A Wu, and Steven P Dow. 2018. CommunityCrit: Inviting the Public to Improve and Evaluate Urban Design Ideas through Micro-Activities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 195.

[76]  Richard F Martell and Cynthia E Willis. 1993. Effects of Observers' Performance Expectations on Behavior Ratings of Work Groups: Memory or Response Bias? *Organizational Behavior and Human Decision Processes* 56, 1 (1993), 91–109.

[77]  David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 224–235.

[78]  Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. 2016. Taking a HIT: Designing around rejection, mistrust, risk, and workers' experiences in Amazon Mechanical Turk. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 2271–2282.

[79]  Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 2243–2251.

[80]  Jennifer Mencl and Douglas R May. 2009. The effects of proximity and empathy on ethical decision-making: An exploratory investigation. *Journal of Business Ethics* 85, 2 (2009), 201–226.

[81]  IG Metall. 2016. Frankfurt paper on platform-based work—Proposals for platform operators, clients, policy makers, workers, and worker organizations. *IG Metall, Frankfurt* (2016).

[82]  Paul Mihas. 2019. Qualitative data analysis. In *Oxford Research Encyclopedia of Education*.

[83] Jeremias Prassl. 2018. *Humans as a service: The promise and perils of work in the gig economy*. Oxford University Press.

[84] Will Qiu, Palo Parigi, and Bruno Abrahao. 2018. More stars or more reviews?. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 153.

[85] Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. 2010. The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, 913–921.

[86] Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. Reputation systems. *Commun. ACM* 43, 12 (2000), 45–48.

[87] Jon Rognerud. 2008. *Ultimate guide to search engine optimization: drive traffic, boost conversion rates and make lots of money*. Jon Rognerud SEO.

[88] Alex Rosenblat and Luke Stark. 2016. Algorithmic labor and information asymmetries: A case study of Uber's drivers. *International Journal of Communication* 10 (2016), 27.

[89] Adam Sadilek, Sean Brennan, Henry Kautz, and Vincent Silenzio. 2013. nemesis: Which restaurants should you avoid today?. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[90] Johannes Saenger, Christian Richthammer, Michael Kunz, Stefan Meier, and Günther Pernul. 2015. Visualizing Unfair Ratings in Online Reputation Systems.

[91] Stefan Schiffner, Sebastian Clauß, and Sandra Steinbrecher. 2011. Privacy, liveliness and fairness for reputation. In *International Conference on Current Trends in Theory and Practice of Computer Science*. Springer, 506–519.

[92] Trebor Scholz. 2017. *Uberworked and underpaid: How workers are disrupting the digital economy*. John Wiley & Sons.

[93] Mirjam Seckler, Alexandre N Tuch, Klaus Opwis, and Javier A Bargas-Avila. 2012. User-friendly locations of error messages in web forms: Put them on the right side of the erroneous input field. *Interacting with Computers* 24, 3 (2012), 107–118.

[94] Aravind Shenoy and Anirudh Prabhu. 2016. Ranking in SEO. In *Introducing SEO*. Springer, 21–35.

[95] MS Siberman and Ellie Harmon. 2017. Rating Working Conditions in Digital Labor Platforms. In *Proceedings of 15th European Conference on Computer-Supported Cooperative Work-Exploratory Papers, Reports of the European Society for Socially Embedded Technologies. http://doi. org/10.18420/ecscw2017-to-be-added.*

[96] M Six Silberman and IG Metall. 2009. Fifteen criteria for a fairer gig economy. *Democratization* 61, 4 (2009), 589–622.

[97] Jiliang Tang, Huiji Gao, Xia Hu, and Huan Liu. 2013. Context-aware review helpfulness rating prediction. In *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 1–8.

[98] Adrián Todolí-Signes. 2017. The end of the subordinate worker? The on-demand economy, the gig economy, and the need for protection for crowdworkers. *International Journal of Comparative Labour Law and Industrial Relations* 33, 2 (2017), 241–268.

[99] C Toxtli, C Flores-Saviaga, M Maurier, A Ribot, T Bankole, A Entrekin, M Cantley, S Singh, S Reddy, and R Reddy. 2018. ExperTwin: An Alter Ego in Cyberspace for Knowledge Workers. In *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. IEEE, 886–891.

[100] Eric M Uslaner. 1999. Trust but verify: Social capital and moral behavior. *Social Science Information* 38, 1 (1999), 29–55.

[101] Andrew Whitby, Audun Jøsang, and Jadwiga Indulska. 2004. Filtering out unfair ratings in bayesian reputation systems. In *Proc. 7th Int. Workshop on Trust in Agent Societies*, Vol. 6. 106–117.

[102] Alex C Williams, Gloria Mark, Kristy Milland, Edward Lank, and Edith Law. 2019. The Perpetual Work Life of Crowdworkers: How Tooling Practices Increase Fragmentation in Crowdwork. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–28.

[103] Li Xiong and Ling Liu. 2003. A reputation-based trust model for peer-to-peer e-commerce communities. In *EEE International Conference on E-Commerce, 2003. CEC 2003.* IEEE, 275–284.

[104] Giorgos Zacharia and Pattie Maes. 2000. Trust management through reputation mechanisms. *Applied Artificial Intelligence* 14, 9 (2000), 881–907.

[105] Runfang Zhou and Kai Hwang. 2007. Powertrust: A robust and scalable reputation system for trusted peer-to-peer computing. *IEEE Transactions on parallel and distributed systems* 18, 4 (2007), 460–473.

[106] Ofer Zwikael, Kazuo Shimizu, and Shlomo Globerson. 2005. Cultural differences in project management capabilities: A field study. *International Journal of Project Management* 23, 6 (2005), 454–462.