TI 2022-086/II
Tinbergen Institute Discussion Paper

Disinformation for Hire: Examining the Production of False COVID-19 Information

Alain Cohn[1]
Jan Stoop[2]
Hatim A. Rahman[3]

1 University of Michigan, School of Information

2 Erasmus University Rotterdam and Tinbergen Institute

3 Northwestern University, Kellogg School of Management

Disinformation for Hire: Examining the Production of False COVID-19 Information

Alain Cohn[*]
Jan Stoop[§]
Hatim A. Rahman[†]

Abstract

Misinformation is linked to increased social divisions and adverse health outcomes. While most research focuses on the spread of misinformation, we examine the production of misinformation intended to mislead (disinformation). Our field experiment (N=1,200) found, adjusting for circumstantial factors, 87% of workers in an online labor market completed a job requesting them to create a misleading COVID-19 graph. Viewing a disinformation graph from the experiment negatively affected people's beliefs and behavioral responses to the COVID-19 pandemic, including increased vaccine hesitancy. Using a "wisdom-of-crowds" approach, we highlight how online labor markets can introduce features that may reduce the production of disinformation.

JEL Codes: C93, D91, J22
Keywords: disinformation, field experiment, online labor markets, immoral work

Acknowledgements: The authors thank participants for their feedback from the Psychology of Technology conference, and seminar participants at Tilburg University, Erasmus University, and University of Michigan. We thank Norina Furrer and Shilei Luo for their excellent research assistance. For financial support, we are thankful to NWO, the Netherlands (Vidi grant VI.Vidi.195.061), University of Michigan, Northwestern University, and Erasmus University. The study was approved by the IRB at the University of Michigan (HUM00179761) and at Erasmus University Rotterdam (IRB-E Approval 2019-06).

[*] University of Michigan, School of Information, adcohn@umich.edu
[§] Erasmus University Rotterdam, stoop@ese.eur.nl
[†] Northwestern University, Kellogg School of Management, hatim.rahman@kellogg.northwestern.edu

1. Introduction

Misinformation continues to proliferate, potentially contributing to a range of adverse outcomes, including swaying political elections (Allcott and Gentzkow 2017, Grinberg et al. 2019), increasing doubts about climate change (Oreskes and Conway 2010), and, most recently, hampering attempts to contain the spread of COVID-19 (Roozenbeek et al. 2020). It is not just scholars sounding the alarm about the pernicious impact of misinformation; a 2021 poll indicated that 95% of Americans believe that misinformation is a problem contributing to social division and unrest. [1]

Recent studies predominantly examine how misinformation is consumed and shared (e.g., Lazer et al. 2018, Vosoughi et al. 2018). While a burgeoning literature focuses on misinformation – information that turns out to be false (Ecker et al. 2022) – less research examines the deliberate production of false information, or what scholars call "disinformation" (van der Linden 2022). Investigating the production of disinformation is crucial considering emerging reports highlighting that a "shadow industry" is proliferating world-wide in which companies and governments hire other firms to create misleading information about their competitors and political opponents, respectively (Bradshaw et al. 2020, Fisher 2021). Although there are anecdotal reports that an industry to produce disinformation is growing, sound empirical evidence is lacking about the extent to which workers are willing to create disinformation for employers.

Studying the production of disinformation is difficult because such unethical work deliberately operates through opacity and obfuscation. Further, identifying the causal effect of a job's ethicality on a person's willingness to complete the job is difficult because one needs to find jobs that are identical (e.g., identical in the type of task, pay, time to complete the job, etc.) except one job asks a person to complete a work request that is considered unethical.

Our pre-registered, IRB approved field experiment overcomes these challenges. We created an employer account and recruited 1,200 workers from one of the largest online labor marketplaces, Amazon Mechanical Turk (MTurk). While most academic research utilize MTurk for surveys, we use the platform to run an experiment where workers complete a regular job and do not know they are part of an experiment (Burbano and Chiles 2021, List and Momeni 2021).[2]

[1] Associated Press-NORC at the University of Chicago Poll: https://apnorc.org/projects/the-american-public-views-the-spread-of-misinformation-as-a-major-problem/, last accessed July 11, 2022.

[2] Although MTurk is known as a platform to conduct surveys and lab style experiments, most jobs are not research related; those jobs that are research related are completed by a small subset of workers (Hauser et al. 2019).

To build credibility with workers, we first hired workers to complete a data visualization job, graphing the number of past COVID-19 infections. Next, to measure the causal effect of a job's ethicality on a person's willingness to accept and complete the job, we offered each worker a new job. This allowed us to randomly invite each worker to a job that either requested them to accurately graph the number of COVID-19 deaths (control condition) or fabricate the number of COVID-19 deaths to make the curve in the graph look flatter compared to the official data (disinformation treatment). To emphasize the potential downstream consequences of completing the job, workers in both treatments were told, before they accepted the job, that the graph they created would be posted on social media. To assess the extent to which ethical concerns affect workers' willingness to complete a disinformation job, relative to wages, we also implemented a treatment in which workers were offered half of the wage they were paid for the first job (lower-pay treatment).[3]

We took several steps to minimize the ethical risks of studying the production of disinformation in a field experiment.[4] For example, we ensured that each worker's participation was voluntary by informing them of each job's instructions before they decided to work on a job, including the request to falsify COVID-19 data. Workers could also withdraw from the job at any time (even after they accepted the job), without facing any reputational penalty. Finally, we debriefed each worker when the experiment was completed (providing each worker the opportunity to delete their data).

To assess whether the disinformation job is perceived as more unethical than the control job, we conducted a survey experiment with a different set of workers (n = 692) who had the same level of experience and performance ratings in the online labor market as those from the field experiment ("manipulation check survey"). To reproduce the conditions worker experienced in the field experiment, we used a between-subject design where participants were assigned to view either the control or disinformation job instructions. After reading the instructions, respondents rated the moral acceptability of the job. Workers also answered other questions related to the perceived ethicality of the job, such as the lowest wage they would accept for completing the job.

[3] Note, the payment in the lower-pay treatment was still above the median wage workers earn in this online labor market (Toxtli et al. 2021).

[4] See online appendix A2 for a detailed explanation of the IRB approval, ethics, and risk assessments of our study.

We subsequently conducted a second survey experiment to assess whether viewing the manipulated COVID-19 graphs creates downstream consequences on people's behavior and beliefs ("downstream consequences survey"). This survey recruited a nationally representative sample of US adults (n = 794) on Prolific. Respondents were presented a graph created in the field experiment, either from the disinformation or control condition. We examined their perceived comfort and safety with activities that, at that time (April 2021), posed an increased risk of catching the virus, such as dining in a restaurant or attending an indoor sporting event. We further evaluated whether viewing a manipulated graph influenced their willingness to follow public health guidelines, such as getting the COVID-19 vaccine. Respondents were debriefed at the end of the survey to prevent participants from believing the manipulated graph was displaying official data.

A third survey experiment explored interventions that may slow the production of disinformation in online labor markets ("platform intervention survey"). We probed the effectiveness of these interventions by leveraging the "wisdom-of-crowds" effect and the fact that forecasters with contextual expertise can provide accurate predictions of the relative effectiveness of interventions (DellaVigna and Pope 2018). The survey recruited experienced workers (n = 400) and focused on five types of platform interventions, including (i) a training video, (ii) a reminder of the terms of service, (iii) incentives for whistleblowing, (iv) a social information nudge, and (v) worker accountability. For each intervention, respondents had to predict how many workers (out of 100) would complete the disinformation job from our field experiment.

In our field experiment we found that 61% of workers completed the second job that asked them to make the COVID-19 death rate look less worrying, which is less than the 70% completion rate in the control condition. Accounting for circumstantial factors, as captured by the control condition, this implies that only a small fraction of workers (13%) declined the disinformation job due to ethical concerns.[5] In the lower-pay treatment, 51% of workers completed the second job. Thus, a 50% decrease in wages is associated with a 27% reduction in the job completion rate, which implies a job-completion price elasticity of 0.54. In other words, the response in labor supply to the disinformation job is similar to reducing wages by 25%. We further found workers – across all treatments – made a concerted effort to complete the job, suggesting that workers were attentive

[5] Since workers were randomly assigned to treatments, we can assume that 30% of workers (1-0.70) across all treatments did not complete the second job for extraneous reasons (e.g., time constraints). Applying this number to the disinformation treatment suggests 13% of workers (1-(0.61/0.70)) declined the disinformation job due to ethical reasons.

to the job instructions and that the high completion rate in the disinformation treatment was not just due to workers' inattentiveness. The high job completion rate is also notable considering that a majority of the respondents in the manipulation check survey indicated that the disinformation job is 'morally unacceptable.'

Our downstream consequences survey suggests that even a single exposure to the COVID-19 disinformation created in the field experiment can contribute to behaviors that put individuals and communities at greater risk to catch the virus. The results show participants are more willing to engage in risky behaviors in the early phase of the pandemic (when the COVID-19 vaccine was not widely available), after seeing the manipulated graph, such as attending an indoor mass-gathering event or dining-in at a restaurant. Participants who saw the manipulated graph were also less worried about the pandemic (e.g., new virus variants) and less likely to express a desire to receive a COVID-19 vaccine. These results highlight the importance of limiting the production of disinformation that could spread on social media.

As such, our last study examines measures that may inhibit the production of disinformation in online labor markets. Our expert forecasters predicted that increasing worker accountability (i.e., signaling greater responsibility when taking on jobs, by suspending workers' accounts who are found to complete unethical jobs) could substantially lower the acceptance rate for the disinformation job compared to the status quo. Other measures such as whistleblowing incentives and reminding workers about the labor market's agreements not to engage in harmful practices also have the potential to reduce the acceptance rate for the disinformation job compared to the status quo, albeit to a lesser extent than increasing worker accountability. These results suggest potential pathways online labor markets can explore to discourage the production of disinformation.

Our paper makes several contributions. First, to the best of our knowledge, this paper provides the first field evidence examining workers' willingness to complete a disinformation job, opening up a new, much needed research stream on the production side of misinformation. Scholars examining moral issues in labor markets have largely focused on situations where individual employees are responsible for initiating unethical behavior (Zitzewitz 2012), such as managers manipulating firm earnings (Bergstresser et al. 2006), doctors overprescribing antibiotics (Linder et al. 2017), or taxi drivers exploiting naïve passengers by taking longer routes (Balafoutas et al. 2013). Less common, however, are studies examining situations in which

employers are responsible for initiating immoral work, such as tobacco companies hiring workers to create misleading marketing material about smoking's health effects (Proctor 2012) or car manufacturers creating an environment encouraging workers to falsify data on a vehicle's carbon emissions (Pierce and Snyder 2008). Employers offering such jobs often try to avoid external scrutiny, which makes studying workers directly engaging in immoral work difficult. A recent lab experiment and observational study suggest that employers have to offer higher wages to attract people to complete immoral work, and that immoral work attracts people who have lower concerns about the morality of such work (Schneider et al. 2022). Our work extends the emerging literature on immoral work with a field experiment identifying the causal effect of a job's ethicality on a person's willingness to take on a job.

Our paper also extends the literature on misinformation. Research in this domain has primarily investigated factors influencing people's ability to detect misinformation and how to improve people's ability to identify misinformation. As it relates to the former, a main reason people may be susceptible to believing in misinformation is people's inattentiveness to the information they are consuming (Pennycook et al. 2021, Pennycook and Rand 2019b). As it relates to the latter, scholars propose two general approaches to increase people's ability to discern between true and false information: "pre-bunking" (e.g., training) and "debunking" (e.g., warning label) (van der Linden 2022). Emerging studies also examine how people's exposure to misinformation can have negative downstream consequences. Bursztyn et al. (2022), for example, show that repeated exposure to misinformation broadcasted on popular opinion television shows is associated with greater numbers of COVID-19 cases and deaths. We contribute to this literature by showing that even a single exposure to misleading graphical information can have negative downstream consequences. The fact that the misleading graph in our study had relatively large effects on people's beliefs and behavioral intentions highlights the importance of further investigating how the frequency (e.g., repeat vs. single), medium (e.g., television vs social media), and type (e.g., text vs. visual) of exposure to misinformation contributes to downstream behaviors.

Third, researchers have highlighted how online labor markets provide access to a large, global labor pool that can be used by employers to increase innovation (Kittur et al. 2013) and complete complex projects more quickly (Valentine et al. 2017). Further, employers benefit from sourcing labor from online labor markets in part due to the greater market power they have on these platforms compared to more traditional labor markets (Dube et al. 2020). Due to the benefits

of online labor markets, economists have focused on interventions which make them more efficient, particularly by making it easier for employers and workers to find each other. Horton (2019), for example, demonstrates how displaying workers' capacity to take on new projects led to increased market surplus by up to 6% due to an increase in the number of employer and worker matches. As another illustrative example, Pallais (2014) highlights how experienced workers receiving detailed feedback (i.e., specific, objective information) from employers about their performance helps these workers secure more jobs and higher wages compared to workers who receive coarse feedback (i.e., uninformative, generic information about their performance). While the literature has largely focused on the efficiency of online labor markets, our study highlights ethical issues, which appear particularly prominently on these labor platforms, and provides insight into how organizations providing these labor platforms may take steps to address these issues.

The remainder of this paper proceeds as follows. Section 2 discusses the experimental designs and procedures. Section 3 presents the experimental results. Finally, section 4 concludes.

2. Experimental designs and procedures

The research was approved by University of Michigan's and Erasmus University's Institutional Review Boards (see online appendix section 2 for a detailed explanation of the IRB approval, ethics, and risk assessments of our study). We preregistered our hypotheses, primary analyses, and sample sizes for each study (non-preregistered analyses are indicated as being post-hoc). Preregistration and data are available online (https://www.socialscienceregistry.org/trials/7243 and [link]). All target sample sizes were determined based on power calculations (power = 0.8, p = 0.05 two-sided) to detect a small effect size (Cohen's d = 0.2).

Field Experiment

We evaluated several online platforms to conduct our field experiment (see online appendix section 1.1). After careful analysis, we conducted our field experiment on MTurk because compared to the available options it is both theoretically and empirically a suitable setting to study the production of disinformation. Theoretically, MTurk is ideal because it allows us to create identical jobs (e.g., identical in the type of task, pay, time to complete the job, etc.), except one job asks a person to produce disinformation. Empirically, MTurk represents one of the largest

online labor markets corporations and individuals can use to hire workers to complete small jobs (Moss et al. 2020). The combination of these features has led scholars to design field experiments exploring worker behavior on MTurk, including recent work examining worker misconduct and shirking (Burbano and Chiles 2021, List and Momeni 2021).

To assess the extent to which ethical concerns affect workers' willingness to complete a disinformation job, we first created a generic employer account on MTurk. Using this employer account, we posted a data visualization job in February 2021 for 1,200 workers (US residents, 500 or more jobs completed, approval rate of 95% or higher). The job asked workers to graph the number of COVID-19 infections. As described in online appendix section 1.1, workers were required to read data from a table and then use a drag and drop interface to create the graph. Five days after workers completed the first job, we invited them for a second job via email.

The second job required the same skills (i.e., using the same drag and drop interface) as the first job but focused on COVID-19 deaths. We randomly assigned each worker to a second job opportunity that either requested them to accurately graph the number of COVID-19 deaths (control condition) or falsify the number of COVID-19 deaths to make the curve in the graph look flatter compared to the official data (disinformation treatment). Workers were offered the same wage ($1.20) for the first and second job, respectively (all payments reported are in US dollars). We code workers' completion of the second job as 1, and 0 otherwise. We compute the number (and percent) of under-reported COVID-19 deaths as the difference between the official data and the data workers plotted.

To assess the extent to which ethical concerns affect workers' willingness to complete a disinformation job, relative to wages, we also implemented a lower-pay condition in which workers were offered half of the wage they were paid for the first job. To highlight the potential impact the created graph could have on others, workers in each treatment were told, before accepting the job, that the graph would be posted on Facebook and Twitter.[6]

We preregistered a sample size of 1,200 US workers to complete the first job, allowing us to assign 400 to each treatment (control, lower-pay, disinformation). 1,197 were retained for the second job (see online appendix section 1.1). Note, we did not collect any background information from the workers because we did not want workers to know they were part of an experiment (until

[6] We planned to post the graphs online, as indicated in the job description. However, IRB required us not to post the graphs on social media to minimize the potential of any harmful consequences (e.g., we cannot control how others online would use the manipulated graphs).

debriefing). For both jobs, we chose official COVID-19 data (www.covidtracking.com) from California because the rise in infection and death rates were among the highest in the US between November 2020 and January 2021.

Manipulation Check Survey

We conducted the manipulation check survey with MTurk workers in May 2021 using Qualtrics. We preregistered a sample size of 800 US workers (400 in each treatment). 692 completed the survey (339 in control, 353 in disinformation). The participants had the same qualifications as the workers from the field experiment (US residents, 500 or more jobs completed, approval rate of 95% or higher). They were paid $1 for completing the survey. Using a between-subject design, subjects were randomly assigned to receive information either about the job in the control or disinformation treatment from our field experiment. If a respondent successfully completed the attention check, they were first shown the same instructions the workers were given in the field experiment to complete the job. Importantly, we asked workers to imagine receiving the job request, without telling them that the job was part of an experiment. Subsequently, respondents were asked to evaluate the respective job's ethically by answering the question: "Do you personally believe that working on this follow-up HIT is morally acceptable, morally unacceptable, or is it not a moral issue?". Respondents could answer that they found the job either "Morally acceptable," "Morally unacceptable," or "Not a moral issue". As pre-registered, we dichotomized and coded participants' ethicality assessment "morally unacceptable" as 1, and "morally acceptable" and "not a moral issue" as 0. Additionally, respondents were asked whether they themselves would accept and complete the job in the respective condition, to predict the number of workers (out of 100) who would accept and work on the job, and to state the lowest wage they would accept to work on the job. Because answers could be influenced by social desirability bias, respondents also filled out the Impression Management subscale (bias toward pleasing others) of the Balanced Inventory of Desirable Responding short form (BIDR-16) (Hart et al. 2015) (see online appendix section 4.2 for survey questions).

Downstream Consequences Survey

We conducted the downstream consequences survey on Prolific in April 2021 using Qualtrics. A nationally representative sample of 794 (800 pre-registered) US residents completed the survey (393 in control, 401 in disinformation). Respondents were paid $0.87 for completing the survey. Using a between-subject design, respondents were randomly assigned to view a graph showing the official data or a graph created in the disinformation treatment. If a respondent successfully completed the attention check, they were asked to imagine viewing the graph on social media (e.g., Facebook or Twitter). Note, they were not told that the graph was from the field experiment or who posted the graph on social media at this point (they were debriefed after completing the survey).[7] After viewing the graph, respondents were asked questions that allow us to better understand the downstream consequences of viewing such a graph on social media, particularly the downstream consequences related to COVID-19 risk perceptions and behavioral intentions. We focused on risk perceptions and behaviors the CDC deemed risky during the COVID-19 pandemic (see online appendix section 4.3 for survey questions). We used 7-point scales to measure risk perceptions and behaviors.

Platform Intervention Survey

For the platform intervention survey, which was conducted in January 2022 using Qualtrics, we recruited MTurk workers as forecasters to leverage their contextual experience and expertise. We preregistered a sample size of 400 US workers; 400 completed the survey. The participants had the same qualifications as the workers from the field experiment and manipulation check (US residents, 500 or more jobs completed, approval rate of 95% or higher). They were paid $0.75 for completing the survey. If a respondent successfully completed the attention check, they were shown the instructions that workers in the disinformation treatment of the field experiment received. They were then asked to predict the number of workers (out of 100) who would accept and work on the disinformation job. Next, to assess the effectiveness of interventions in deterring workers from accepting and completing the disinformation job, forecasters were shown five platform interventions, including (i) a training video showing examples of jobs violating the platform's terms of service, (ii) a reminder of the platform's terms of service condition that prohibits engaging in harmful activities, (iii) incentives for whistleblowing, (iv) a social information nudge that allows a worker to view how many other workers declined a job, and (v) a

[7] We did not include a source for the data to recreate the conditions people encounter on social media; information on social media is often not attributed to a fact-checked source (Pennycook and Rand 2019a).

worker accountability policy that would lead to account suspension for engaging in unethical jobs. Using a within-subjects design, we randomized the order of the interventions for each forecaster (see online appendix section 1.4 for the theoretical motivation of the interventions). Forecasters predictions were measured on a scale between 0 to 100.

3. Experimental Results

As preregistered, for each study, we use nonparametric tests (Fisher's exact tests for binary variables and Mann Whitney rank-sum tests for discrete variables) and OLS regressions. The regressions allow us to control for background characteristics. To correct for multiple hypothesis testing, we also report Bonferrroni adjusted p-values (see online appendix section 3). For the within-subjects analysis of the platform intervention survey, we use pairwise Wilcoxon signed-rank tests.

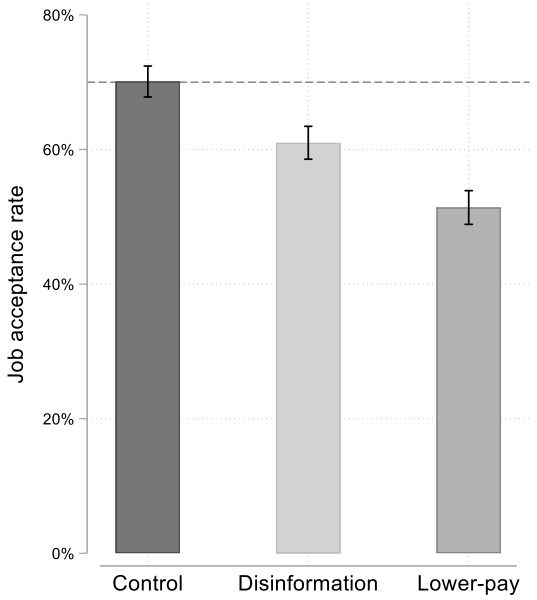
Results of the Manipulation Check Survey

We first examine whether workers find it unethical to work on a job asking them to manipulate COVID-19 data that would be posted on social media. First, about one in ten respondents (10.3%) reported they have previously encountered jobs asking them to fabricate or manipulate data in a misleading way. The question likely captures only a subset of unethical jobs posted on MTurk, suggesting that workers encounter unethical jobs similar to the one we offered. Second, 44.8% of the respondents who were presented with the instructions for the disinformation job said the job is 'morally unacceptable' compared to only 6.0% who were shown the control job ($p < 0.001$). Respondents in the disinformation condition further indicated that they would require more than 30% higher compensation ($2.38 vs. $1.81) for completing the job ($p < 0.001$), and that they would be less likely to accept and complete the job (62.6% vs 88.8%; $p < 0.001$). Additional regression analyses suggest that the results are not just driven by social desirability considerations (e.g., to conform to social expectations). Restricting the sample to respondents with a lower propensity to provide socially desirable answers (based on a median split of the Impression Management score) yields qualitatively similar results to the full sample (see online appendix Table A10). Overall, the results suggest that our experimental manipulation worked as intended.

Results of the Field Experiment

Turning to the field experiment, we assess the actual willingness to falsify COVID-19 data. Figure 1 shows that 61% of the workers completed the job that asked them to make the COVID-19 death rate look less worrying. In the control condition, 70% of the workers completed the job, which is significantly more than in the disinformation treatment (p = 0.007). Since we randomly assigned workers to conditions, this implies that, across conditions, about a third of the workers did not work on the job for reasons unrelated to ethical concerns (e.g., lack of time). When accounting for these circumstantial factors, an estimated 87% (i.e., 61%/70%) of workers were willing to falsify COVID-19 data that would be shared on social media.

Figure 1: Job acceptance rates by treatment



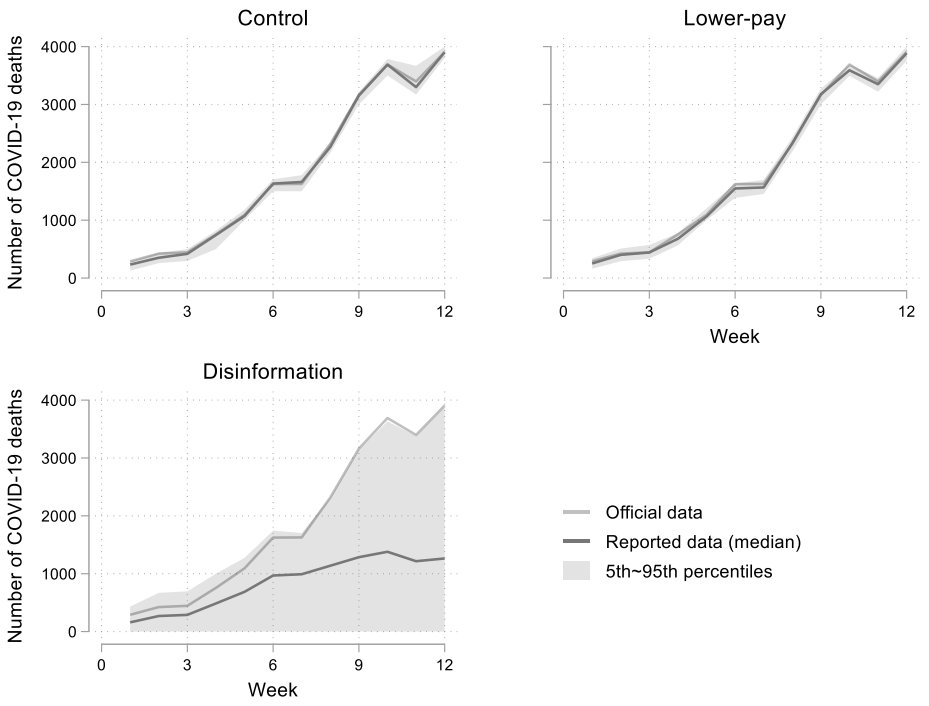
Notes: This figure shows the percentage of workers who completed the second job by treatment. Error bars indicate s.e.m.

The lower-pay condition enables us to assess the extent to which ethical concerns affect labor supply, relative to wages. We find that 51% of the workers completed the job when they were offered half the wage they earned for the first job. If we again take into account circumstantial reasons for not doing the job, the completion rate amounts to 73% in the lower-pay condition. Thus, as one would expect, labor supply decreases significantly with lower wages (p < 0.001). Using a back-of-the-envelope calculation, we estimate the effect of the disinformation treatment to be approximately the same as paying workers 25% less.

Unlike in the control and lower-pay conditions, workers in the disinformation treatment were asked to make the COVID-19 death data look less worrying, without specifying by how much they needed to reduce the death numbers. Figure 2 shows, by treatment, the median graph that the workers produced along with the 5th and 95th percentiles. As evident from the figure, the trend in COVID-19 deaths looks less worrisome in the disinformation treatment compared to the other two conditions. On average, workers in the disinformation treatment under-reported 11,814 deaths, which represents a 52.0% reduction in deaths reported compared to the official data. The number of deaths reported is also significantly different from the control condition ($p < 0.001$). Unlike the disinformation treatment workers in the lower-pay treatment produced a graph that matches the official data to a similar extent as the control condition. In the control condition, workers under-reported deaths by 1.4% on average, whereas in the lower-pay treatment, they overreported by deaths by 1.9% ($p < 0.001$). This suggests that even when paid substantially less than the first job (and the control condition), workers still took the job seriously.

Figure A1 in the online appendix displays the cumulative distribution function of the total number of under-reported COVID-19 deaths for each condition. The figure shows that it was not just a few workers who reduced the number of deaths in the disinformation treatment. For example, 80% of them under-reported the number of COVID-19 fatalities by at least 7,000 deaths compared to less than 1% of workers in the control condition. A Kolmogorov-Smirnov test confirms that the two distributions are statistically different ($p < 0.001$). Together, this not only suggests that a majority of the workers were willing to work on the disinformation job, but also that they considerably reduced COVID-19 deaths compared to the official data.

Figure 2: Number of COVID-19 deaths reported by treatment

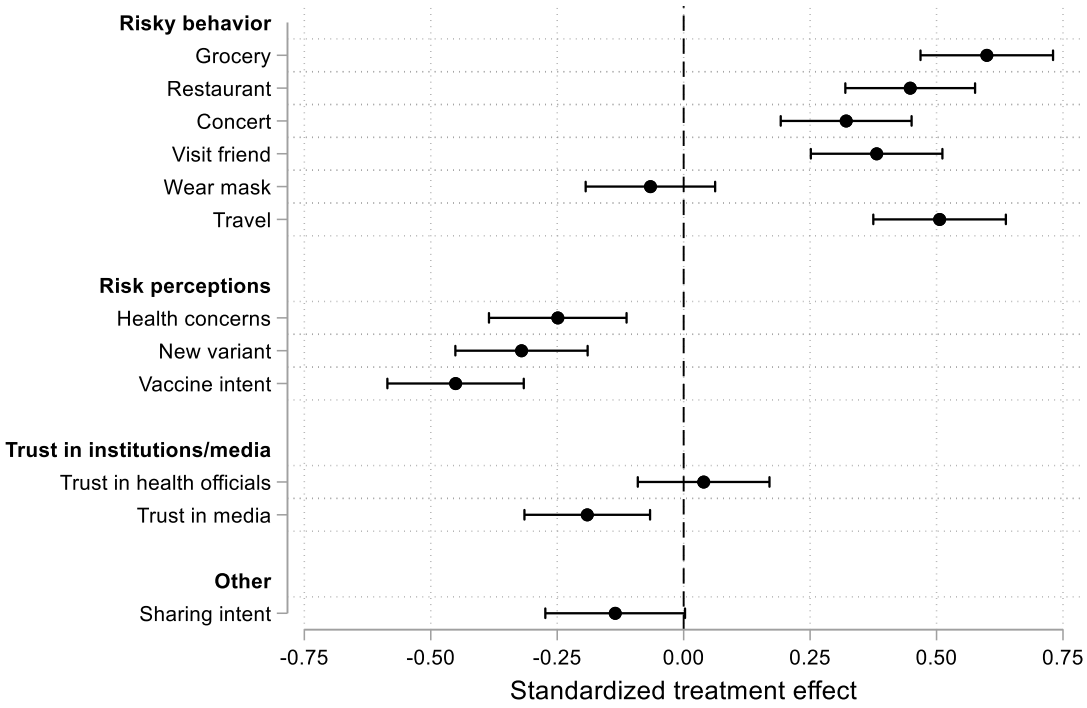


Notes: This figure shows for each week in the dataset the median number (darkest gray line) of COVID-19 deaths reported in each treatment. The shaded area shows the 5th and 95th percentiles.

Results of the Downstream Consequences Survey

Do the graphs created in the field experiment have the potential to affect people's beliefs and behavioral responses to the pandemic? We answered this question by conducting a survey experiment with a nationally representative sample examining the downstream effects of the manipulated COVID-19 graphs. To more easily interpret and compare the results, we standardize the outcomes and summarize the regression results in Figure 3.

Figure 3: Effect of manipulated COVID-19 graph on beliefs and behavioral intentions



Notes: This figure shows standardized effect sizes from regressions of the manipulated COVID-19 graph on risky behaviors, risk perceptions, trust in institutions/media, and sharing intentions. Error bars indicate s.e.m.

Overall, the figure suggests that respondents were more willing to engage in activities that could increase their exposure to COVID-19 after seeing a graph created in the disinformation treatment. Respondents who were presented with the manipulated graph said they would feel more comfortable going to the grocery store (0.60 SD, 95% CI = [0.47, 0.73], $p < 0.001$), eating out in a restaurant (0.45 SD, 95% CI = [0.32, 0.58], $p < 0.001$), attending an indoor sporting event or concert (0.32 SD, 95% CI = [0.19, 0.45], $p < 0.001$), visiting a close friend or family member inside their home (0.38 SD, 95% CI = [0.25, 0.51], $p < 0.001$), and travel in the next month (0.51 SD, 95% CI = [0.37, 0.64], $p < 0.001$). The only item where we do not observe a stark difference between conditions is when we asked respondents about their support for mandatory mask wearing in public places (-0.07 SD, 95% CI = [-0.19, 0.06], $p = 0.314$). The support for a mask mandate was generally high, possibly due to the fact they were already in place in many states at the time of the survey.

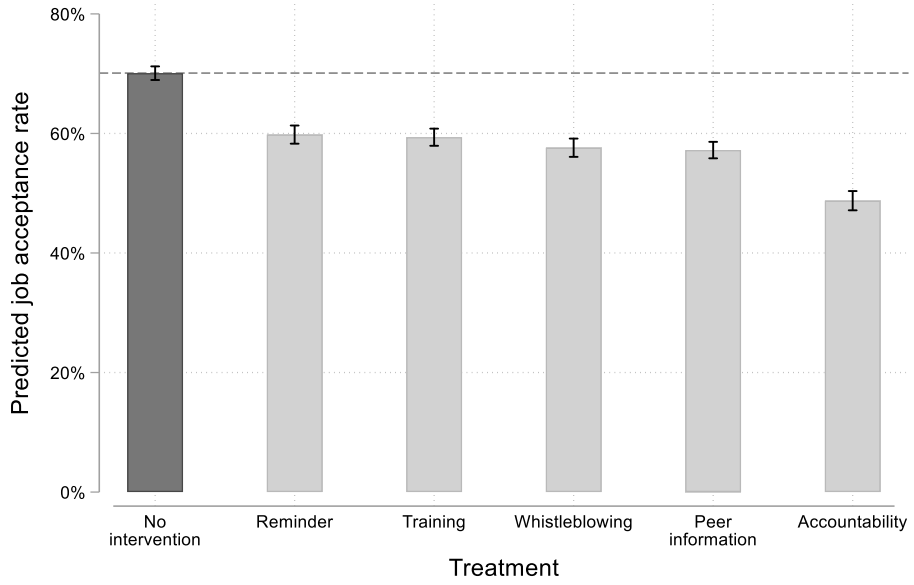
The results look similar for COVID-19 risk perceptions. Those exposed to the manipulated graph were less worried about COVID-19 health consequences (-0.25 SD, 95% CI = [-0.38, -0.11],

$p < 0.001$), that the new variant ("Delta") at that time would lead to a new wave of infections (-0.32 SD, 95% CI = [-0.45, -0.19], $p < 0.001$), and they expressed a lower willingness to get the COVID-19 vaccine (-0.45 SD, 95% CI = [-0.59, -0.32], $p < 0.001$). These results may actually underestimate the downstream consequences of people viewing manipulated COVID-19 graphs because the questions do not factor in the possibility that people may spread the disinformation. Although respondents tended to indicate they would be more likely to fact check the graph (0.09 SD, 95% CI = [-0.05, 0.23], $p = 0.188$, post-hoc), reproducing what we would expect people do on social media, they were only slightly less likely to say they would share the graph with their friends and followers on social media (-0.14 SD, 95% CI = [-0.27, -0.03], $p = 0.055$). More politically conservative respondents, however, were significantly more likely to share the manipulated graph compared to the control graph (0.09 SD, 95% CI = [0.01, 0.16], $p = 0.034$). Exposure to COVID-19 disinformation could also lower trust in the mainstream media and public health officials (e.g., CDC). Respondents who viewed the manipulated graph had less trust in the mainstream media (-0.19 SD, 95% CI = [-0.31, -0.07], $p = 0.003$). However, there was no effect on their trust in how public health officials were managing the COVID-19 outbreak (0.04 SD, 95% CI = [-0.09, 0.17], $p = 0.551$). Overall, the results highlight that even a single exposure to COVID-19 disinformation can contribute to downstream behaviors that put individuals and communities at greater risk to catch the virus.

Results of the Platform Intervention Survey

Given the large share of workers who completed the disinformation job, we tested five interventions that may deter workers from accepting such jobs. We provided a new set of workers ("forecasters") with the instructions for the disinformation job and then asked them to predict how many workers (out of 100) would be willing to complete the job. To obtain a benchmark estimate, forecasters first predicted the acceptance rate for the disinformation job if no intervention was introduced. Shown in Figure 4, they predicted, on average, that about 70 out of 100 workers will accept to work on the disinformation job, which is in between the unadjusted completion rate (61%) and the rate adjusted for circumstantial factors (87%). This indicates that the forecasters were well-calibrated.

Figure 4: Effect of platform interventions on predicted job acceptance rates



Notes: This figure shows forecasters' predictions of the number of workers who would accept the disinformation job in response to each platform intervention. Error bars indicate s.e.m.

Turning to the interventions, forecasters thought that expanding the platform's terms of service – such that not only the employer's but also the worker's account would be suspended for working on a job violating the terms of service – would be most effective in deterring workers from accepting the disinformation job. On average, they predicted that 49 out of 100 workers would work on the job if the accountability intervention was implemented, which is a significantly lower acceptance rate compared to the status quo ($p < 0.001$). All the other interventions were also predicted to significantly deter workers from accepting the disinformation job, though to a lesser extent than the accountability condition. For example, forecasters predicted that 58 out of 100 workers would work on the disinformation job if the platform implemented a financial incentive for whistleblowing, which would offer workers 10% of what the job pays. Forecasters predicted a similar acceptance rate (57 out of 100) for an intervention providing workers with information about how many times a job has been viewed by other workers (instead of only showing how many times the job has been completed). This information would provide workers a sense for how many people decided not to work on a job. Forecasters also predicted a similar acceptance rate for the

remaining two interventions – a training video showing examples of jobs that violate the platform's terms of service and a rule requiring all workers to reaffirm, before accepting a new job, that they agree to the current terms of service prohibiting engagement in harmful activities. All of these interventions have a significantly lower predicted acceptance rate than the status quo ($p < 0.001$), yet were not considered as effective as the worker accountability condition ($p < 0.001$).

4. Conclusion

We conducted a field experiment examining workers' willingness to complete a disinformation job in an online labor market. Accounting for circumstantial factors (e.g., lack of time), we found 87% of workers completed the job requesting them to manipulate a COVID-19 graph by falsifying the underlying numbers. These workers accepted the disinformation job even though the instructions stated that the graph they created would be shared publicly on social media and the manipulation check survey showing the disinformation job is perceived as unethical.

The willingness of so many workers to complete the disinformation job is surprising in light of the literature suggesting that people often behave ethically even when faced with considerable financial incentives to behave unethically and there is little chance of others discovering their unethical actions (Abeler et al. 2019, Cohn et al. 2019). Relatedly, research using MTurk workers as subjects do not find them to be particularly honest or dishonest compared to other populations (Peer et al. 2022, Snowberg and Yariv 2021). Further, the lower-pay condition shows that MTurk workers do not only care about earning money; many workers in the lower-pay treatment were not willing to complete the second job, despite workers being offered more than the median wage on this platform (Toxtli et al. 2021).

While our study does not provide direct evidence for why so many workers chose to complete the disinformation job, the literature on markets and morality points to market features conducive to unethical behavior. For example, scholars suggest competition can contribute to a person's willingness to behave unethically (e.g., Shleifer 2004). This raises the possibility that in highly competitive labor market contexts, workers may be more willing to take on an unethical job because they may believe someone else would complete it if they do not. Another potential factor is the anonymity between market participants, which may reduce perceptions of responsibility (e.g., Kirchler et al. 2016). On MTurk, for example, both employers and workers are largely anonymous, potentially diffusing responsibility over the downstream consequences of

a job. Thus, future work should investigate the extent to which competition for jobs, anonymity, and other factors salient in online labor markets may contribute to unethical behavior, such as the production of disinformation.

Our downstream consequences survey suggests that exposure to misleading graphical information can have negative effects on people's health related perceptions and behaviors, such as vaccine hesitancy. While we measured participants' behavioral intentions, rather than the actual behaviors, recent research suggests that behavioral intentions are predictive of actual behavior, including the propensity to receive a COVID-19 vaccine (e.g., Campos-Mercade et al. 2021). Thus, the fact that the misleading graphs in our study had large effects, relative to other studies that examine misinformation communicated through text (e.g., Greene and Murphy 2021), highlights the importance of investigating the persuasiveness of misinformation across different communication modalities. This is particularly important given the increased popularity of video-based social media platforms (e.g., TikTok).

While there are currently few guardrails in place to prevent the production of disinformation in online labor markets, the results of our platform intervention survey suggest steps platforms may take to inhibit the production of disinformation. Simple interventions, such as reminding workers that accepting unethical jobs violate the platform's terms of service, may reduce the likelihood that people accept jobs requesting them to create disinformation. Additionally, while many online labor platforms have mechanisms for workers to report unethical job requests, our study suggests that platforms should experiment with offering incentives for accurately reporting such jobs. In our context, for example, the platform did not provide any incentive for a worker to report an unethical job. Currently, workers have to be willing to sacrifice their pay when reporting such a job. As a result, for workers, it is financially more beneficial to work on a job, regardless of its ethicality, compared to reporting it for violating the terms of service. Our study highlights the importance of not only focusing on interventions to address misinformation's spread on social media, but also on changing features of the market environment that can hinder the production of disinformation in online labor markets, before it has a chance to spread.

In conclusion, our study highlights the ease in which virtually anyone can hire people to produce disinformation, that the disinformation can have negative downstream consequences, and that there may be simple interventions to slow the production of disinformation in online labor

markets. Additional research on the production of disinformation is particularly important because scholars warn that new methods to create disinformation are being developed, including by AI systems (Köbis et al. 2021).

References

Abeler J, Nosenzo D, Raymond C (2019) Preferences for truth-telling. *Econometrica* 87(4):1115–1153.

Allcott H, Gentzkow M (2017) Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31(2):211–236.

Balafoutas L, Beck A, Kerschbamer R, Sutter M (2013) What Drives Taxi Drivers? A Field Experiment on Fraud in a Market for Credence Goods. *The Review of Economic Studies* 80(3):876–891.

Bergstresser D, Desai M, Rauh J (2006) Earnings Manipulation, Pension Assumptions, and Managerial Investment Decisions. *The Quarterly Journal of Economics* 121(1):157–195.

Bradshaw S, Bailey H, Howard P (2020) *Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation* (Programme on Democracy & Technology, Oxford, UK).

Burbano VC, Chiles B (2021) Mitigating Gig and Remote Worker Misconduct: Evidence from a Real Effort Experiment. *Organization Science*.

Campos-Mercade P, Meier AN, Schneider FH, Meier S, Pope D, Wengström E (2021) Monetary incentives increase COVID-19 vaccinations. *Science* 374(6569):879–882.

Cohn A, Maréchal MA, Tannenbaum D, Zünd CL (2019) Civic honesty around the globe. *Science* 365(6448):70–73.

DellaVigna S, Pope D (2018) Predicting Experimental Results: Who Knows What? *Journal of Political Economy* 126(6):2410–2456.

Dube A, Jacobs J, Naidu S, Suri S (2020) Monopsony in Online Labor Markets. *American Economic Review: Insights* 2(1):33–46.

Ecker UKH, Lewandowsky S, Cook J, Schmid P, Fazio LK, Brashier N, Kendeou P, Vraga EK, Amazeen MA (2022) The psychological drivers of misinformation belief and its resistance to correction. *Nat Rev Psychol* 1(1):13–29.

Fisher M (2021) Disinformation for Hire, a Shadow Industry, Is Quietly Booming. *The New York Times* (July 25) https://www.nytimes.com/2021/07/25/world/europe/disinformation-social-media.html.

Gray ML, Suri S (2019) *Ghost work: how to stop Silicon Valley from building a new global underclass* (Houghton Mifflin Harcourt, Boston).

Greene CM, Murphy G (2021) Quantifying the effects of fake news on behavior: Evidence from a study of COVID-19 misinformation. *Journal of Experimental Psychology: Applied* 27(4):773–784.

Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D (2019) Fake news on Twitter during the 2016 U.S. presidential election. *Science*.

Hart CM, Ritchie TD, Hepper EG, Gebauer JE (2015) The Balanced Inventory of Desirable Responding Short Form (BIDR-16). *SAGE Open* 5(4):2158244015621113.

Hauser D, Paolacci G, Chandler J (2019) Common Concerns with MTurk as a Participant Pool: Evidence and Solutions. *Handbook of Research Methods in Consumer Psychology*. (Routledge).

Horton JJ (2019) Buyer Uncertainty About Seller Capacity: Causes, Consequences, and a Partial Solution. *Management Science* 65(8):3518–3540.

Kirchler M, Huber J, Stefan M, Sutter M (2016) Market Design and Moral Behavior. *Management Science* 62(9):2615–2625.

Kittur A, Nickerson JV, Bernstein M, Gerber E, Shaw A, Zimmerman J, Lease M, Horton J (2013) The Future of Crowd Work. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. CSCW '13. (ACM, New York, NY, USA), 1301–1318.
Köbis N, Bonnefon JF, Rahwan I (2021) Bad machines corrupt good morals. *Nat Hum Behav* 5(6):679–685.
Lazer DMJ, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, et al. (2018) The science of fake news. *Science* 359(6380):1094–1096.
van der Linden S (2022) Misinformation: susceptibility, spread, and interventions to immunize the public. *Nat Med*.
Linder JA, Meeker D, Fox CR, Friedberg MW, Persell SD, Goldstein NJ, Doctor JN (2017) Effects of Behavioral Interventions on Inappropriate Antibiotic Prescribing in Primary Care 12 Months After Stopping Interventions. *JAMA* 318(14):1391–1392.
List JA, Momeni F (2021) When Corporate Social Responsibility Backfires: Evidence from a Natural Field Experiment. *Management Science* 67(1):8–21.
Moss AJ, Rosenzweig C, Robinson J, Jaffe SN, Litman L (2020) Is it ethical to use Mechanical Turk for behavioral research? Relevant data from a representative survey of MTurk participants and wages.
Oreskes N, Conway EM (2010) Defeating the merchants of doubt. *Nature* 465(7299):686–687.
Pallais A (2014) Inefficient Hiring in Entry-Level Labor Markets. *American Economic Review* 104(11):3565–3599.
Peer E, Rothschild D, Gordon A, Evernden Z, Damer E (2022) Data quality of platforms and panels for online behavioral research. *Behav Res* 54(4):1643–1662.
Pennycook G, Epstein Z, Mosleh M, Arechar AA, Eckles D, Rand DG (2021) Shifting attention to accuracy can reduce misinformation online. *Nature* 592(7855):590–595.
Pennycook G, Rand DG (2019a) Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116(7):2521–2526.
Pennycook G, Rand DG (2019b) Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188:39–50.
Pierce L, Snyder J (2008) Ethical Spillovers in Firms: Evidence from Vehicle Emissions Testing. *Management Science* 54(11):1891–1903.
Proctor RN (2012) *Golden Holocaust: Origins of the Cigarette Catastrophe and the Case for Abolition* 1 edition. (University of California Press, Berkeley).
Roozenbeek J, Schneider CR, Dryhurst S, Kerr J, Freeman ALJ, Recchia G, van der Bles AM, van der Linden S (2020) Susceptibility to misinformation about COVID-19 around the world. *R. Soc. open sci.* 7(10):201199.
Schneider FH, Brun F, Weber RA (2022) Sorting and wage premiums in immoral work. *Working Paper*.
Shleifer A (2004) Does competition destroy ethical behavior? *American economic review* 94(2):414–418.
Snowberg E, Yariv L (2021) Testing the Waters: Behavior across Participant Pools. *American Economic Review* 111(2):687–719.
Toxtli C, Suri S, Savage S (2021) Quantifying the invisible labor in crowd work. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW2):1–26.
Valentine MA, Retelny D, To A, Rahmati N, Doshi T, Bernstein MS (2017) Flash Organizations: Crowdsourcing Complex Work By Structuring Crowds As Organizations.

Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology. UIST '17. (ACM, Denver Colorado USA).
Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. *Science*.
Zitzewitz E (2012) Forensic Economics. *Journal of Economic Literature* 50(3):731–769.

Online Appendix for:

Disinformation for Hire: Examining the Production of False COVID-19 Information

Contents

1. Material and Methods

We conducted one field experiment and three survey experiments. Below, we provide more details about each study's design.

1.1 Field Experiment

Evaluating Options for the Field Experiment: To determine an ideal setting for identifying the causal effect of a job's ethicality on a person's willingness to produce disinformation, we evaluated several platforms, including Upwork, Freelancer, Fiverr, MTurk, ClickWorker, and Proginn (a Chinese online labor platform).

After careful analysis, we conducted our field experiment on MTurk because compared to the available options it is both theoretically and empirically an ideal setting to study the production of disinformation. Theoretically, MTurk is ideal because it allows us to create identical jobs (e.g., identical in the type of task, pay, time to complete the job, etc.), except one job asks a person to produce disinformation. Empirically, MTurk represents one of the largest online labor markets corporations and individuals can use to hire workers to complete small jobs. The combination of these features has led scholars to design field experiments exploring worker behavior on MTurk, including recent work examining worker misconduct and shirking.

Upwork is an online labor market catering towards higher-paying, unstructured jobs such as software engineering, data analysis, and graphic design. We initially registered and ran a pilot study on Upwork. This process showed that the platform is primarily intended for project-based work in which a worker and employer regularly interact. For the purposes of running a field experiment, this arrangement was not ideal because we wanted to keep the communication we sent to each worker consistent. Further, we found the platform was not conducive to hiring a large number of workers at the same time, which was problematic given the study's required sample size. Freelancer.com had a similar setup as Upwork, but we also found that Freelancer does not require users to verify their registration information, reducing the certainty of who is actually being hired.

Fiverr is also an online labor market that allows employers to hire freelancers. We created both a worker and employer account to explore if it was suitable for our field experiment. Similar to Upwork and Freelancer, the platform is primarily set up to hire an individual freelancer to complete a project. We did not find a feature or mechanism on the platform to hire multiple workers to complete a job, making it difficult to implement our field experiment. ClickWorker distinguishes itself from other online labor platforms by finding and hiring freelancers for employers. That is, employers submit a job request and ClickWorker hires people to complete the job; employers are not given additional information about

workers or cannot directly communicate with workers. This setup was not conducive to our field experiment, particularly not being able to randomize workers into different treatments.

Proginn is considered one of China's largest online labor markets for middle and high-skilled software programmers. Similar to Upwork, we found this platform is primarily intended for one-on-one or small team project-based work in which a worker and employer interact regularly. As a result, as detailed in the main text, we conducted our field experiment on MTurk because, compared to the available options, it is both theoretically and empirically an ideal setting to study the production of disinformation using a field experiment approach.

Design of the Field Experiment: To assess the extent to which ethical concerns affect workers' willingness to complete a disinformation job, we first created a generic employer account on MTurk. Using this employer account, we posted a data visualization job in February 2021 for 1,200 workers (US residents, 500 or more jobs completed, approval rate of 95% or higher). The job asked workers to graph the number of COVID-19 infections. As described in section 4.1, workers were required to read data from a table and then use a drag and drop interface to create the graph. Such a job can be done by a human, but it is practically impossible for a bot to complete given the skills required to drag and drop the data points on the graphical interface to match the data in the table. Further, when using the drag and drop interface, workers could see how much their graph differed from the original data, graphically and numerically in a table. For all of the jobs, workers could view a job's title, description, and instructions before deciding whether to work on the job (see section 4.1).

Workers were paid $1.20 for the first and second job and given a maximum of 12 minutes to complete the job. Based on pilot studies, we estimated that it would take them about six minutes to complete the job. Accordingly, this wage equates to paying a worker $12 per hour, which is in line with what the average employer pays on MTurk ($11, Hara et al. 2018) and within the range of minimum wages offered in the US (https://www.dol.gov/agencies/whd/minimum-wage/state).

This constituted the first stage of our field experiment. We needed this first job because MTurk does not provide employers with information about how many people decline a job offer. It also helped us to make sure workers were able to perform the task and it created some familiarity with the employer.

MTurk provides employers with an Application Programming Interface (API) that allows employers to send a custom email message to workers who employers have hired previously. As such, five days after workers completed the first job, we invited them for a second job via the API. The second job required the same skills (i.e., using the same drag and drop interface) as the first job but focused on COVID-19 deaths. It was kept open for seven days to ensure that workers had sufficient time to notice the job invitation. The email invitation for the second job allowed us to randomly assign each worker we previously

hired to a new job that either requested them to accurately graph the number of COVID-19 deaths (control condition) or falsify the number of COVID-19 deaths to make the curve in the graph look flatter compared to the official data (disinformation treatment). Although workers in the disinformation treatment did not receive detailed instructions on how much to reduce the number of deaths, they were told to especially reduce the number of deaths after a specific date so that the curve would look flatter and thus less worrying (see section 4.1). The control condition gives us an estimate of how many workers turned down the disinformation job because of reasons unrelated to their ethical concerns.

To get a sense for the magnitude of workers' willingness to complete a disinformation job, we also randomly assigned one third of the workers to a lower-pay condition in which workers were offered half of the wage ($0.60) that they were previously paid for in the first job. Although workers in the lower-pay treatment were offered considerably less than what they were paid for the first job, they still earned more than the median wage workers earn on MTurk (Toxtli et al. 2021).

As shown in section 4.1, in all of the treatments, the second job's invitation and instruction messages included the following sentence, "I will publicly post this new graph on Facebook and Twitter." We included this sentence to enhance workers' perception that their decision to complete the job could influence the broader public's beliefs and behaviors related to COVID-19. In other words, we wanted workers to consider the potential real-world, downstream consequences of their actions when deciding whether to work on the job. As required by IRB, to minimize the potential of any unintended consequences, we did not post the graphs on social media and debriefed the participants.

For the field experiment, we pre-registered 1,200 participants (i.e., 400 per condition) based on a power calculation. According to our power calculation, we needed at least 394 participants for each of the three treatments to detect a small effect (Cohen's $d = 0.2$) on job completion with a power of 0.8 and significance threshold of 0.05 (two-sided). As such, we allowed 1,200 workers to work on the first job so that we could assign 400 to each treatment for the second job. We had incomplete data for two workers. One worker submitted the first job twice (this was possible because we posted the first job in batches); we kept the worker's first submission. Thus, our final sample has 1,197 workers (398 workers in the control treatment, 399 workers in the lower-pay treatment, and 400 workers in the disinformation treatment). Because of the nature of our field experiment (i.e., posing as a real employer), we did not ask about workers' background characteristics and can therefore not check for balance across treatments.

Workers who encounter unusual job requests can post them to discussion boards to alert other workers that they should avoid certain employers and jobs. Given the nature of the disinformation job, we monitored several popular MTurk discussion boards (MTurkForum, MTurkCrowd, TurkverView, Turkopticon, and Reddit). During the duration of the field experiment, we observed one worker reporting that they received the disinformation job invitation. Another worker replied to the message, unsure if the

request was real, and indicated they will not work on the job due to its ethicality. Three workers contacted us via the employer account expressing reservations about completing the job. These workers are retained in the dataset; two workers declined the disinformation job for ethcial reasons and the third worker was retained because they contacted us after completing the job.

1.2. Manipulation Check Survey Design Details

We conducted the manipulation check survey with MTurk workers in May 2021 using Qualtrics. The participants had the same qualifications as the workers from the field experiment (US residents, 500 or more jobs completed, approval rate of 95% or higher). They were paid $1 for completing the survey, which was expected to take four minutes or less (median completion time was 3 minutes and 45 seconds). Using a between-subject design, subjects were randomly assigned to receive information either about the job in the control or the disinformation treatment from our field experiment. If a respondent successfully completed the attention check, they were first shown the same instructions that the workers were given in the field experiment to complete the job. Importantly, we asked workers to imagine receiving the job request, without telling them that the job was part of an experiment.

Subsequently, respondents were asked to evaluate the respective job's ethically by answering the question: "Do you personally believe that working on this follow-up HIT is morally acceptable, morally unacceptable, or is it not a moral issue?". Respondents could answer that they found the job either "Morally acceptable," "Morally unacceptable," or "Not a moral issue". This question serves as a manipulation check to determine whether the disinformation job was perceived as more unethical than the control job. As pre-registered, we combined the answer options "morally acceptable" and "not a moral issue" so that we could identify in each condition the share of workers who found the job unethical. To compare people's self-reported behavior with workers' actual behavior in our field experiment, we asked respondents whether they themselves would accept and complete the job in the respective condition. Because of potential social desirability bias, we further asked them to predict the number of workers (out of 100) who would accept and work on the job. We also asked them to state the lowest wage they would accept to work on the job using a sliding scale between $0 and $5. Assuming that people would need to be compensated for working on an unethical job, this serves as another indicator of people's perceived ethicality of the job. To understand whether workers felt responsible for how the graph they made could affect people who saw it on social media, we also asked them about perceived responsibility on a 7-point scale from 1 ("Not at all responsible") to 7 ("Fully responsible").

To get a sense of the prevalence of unethical jobs involving data fabrication or manipulation, respondents were asked whether they ever encountered such jobs on MTurk. If so, respondents were provided a free response field to describe the type of disinformation jobs they encountered. Because answers to these questions could be impacted by social desirability bias, respondents filled out the Impression Management subscale (bias toward pleasing others) of the Balanced Inventory of Desirable Responding short form questionnaire (BIDR-16) (Hart et al., 2015). A subject is presented with eight statements that are answered on a 7 item Likert scale (1 = "Not true" to 7 = "True"). For each statement for which a subject responds with one of the two most extreme answers (6 or 7 for normally coded questions,

1 or 2 for reverse coded questions), the subject receives one point. The sum of the points is a subject's impression management score.

Table A1: The BIDR-16 Impression Management Subscale

Please write for each statement how much you agree with it:	1 = "Not true" 7 = "True"
1. "I sometimes tell lies if I have to." (r)	1-2
2. "I never cover up my mistakes.	6-7
3. "There have been occasions where I have taken advantage of someone." (r)	1-2
4. "I sometimes try to get even rather than to forgive and forget." (r)	1-2
5. "I have said something bad about a friend behind his or her back." (r)	1-2
6. "When I hear people talk privately, I avoid listening."	6-7
7. "I never take things that don't belong to me."	6-7
8. "I don't gossip about other people's business."	6-7

Notes: This table shows the questions used to assess subjects' susceptibility to give socially desirable answers, based on the BIDR-16 Impression Management subscale. For each statement, if a subject responds with one of the two extreme answers (6 or 7 for normally coded questions, 1 or 2 for reverse-coded (r) questions), the subject receives one point. The sum of the points is a subject's impression management score.

Finally, we asked respondents several background demographic and political orientation questions. The complete survey can be found in section 4.2.

For this survey, we pre-registered 800 participants (i.e., 400 per condition) based on a power calculation. According to our power calculation, we needed at least 394 participants in each condition to detect a small effect (Cohen's d = 0.2) for the perceived ethicality question (binary variable) with a power of 0.8 and significance threshold of 0.05 (two-sided). 1,107 MTurk workers started the survey, 892

successfully passed the attention check, 828 completed the survey, and 797 successfully submitted the completion code on MTurk.

Note, MTurk provides a unique, anonymized id for every worker, which allows us to track individuals who participated in our field experiment and surveys. This allowed us to create a "blacklist" preventing people from participating in more than one study. This step ensured that participants did not have any prior knowledge that could influence their behavior or responses. Due to an error when updating the "blacklist," however, 105 respondents from the field experiment participated in the manipulation check. These respondents were removed from the analysis (the results are almost identical if these respondents are included). Thus, our final sample has 692 participants. We did not find any other workers participating in subsequent surveys (i.e., platform intervention survey).

A balance check of the background characteristics shows that the variables are balanced across conditions. Additionally, none of the differences across conditions are significant at the 5% level (see Table A2 on the next page).

Table A2: Balance Check for the Manipulation Check Survey

	Control	Disinformation		
	(1)	(2)	(3)	(4)
	Mean (SD)		Δ	p-value
Age	40.664	42.390	-1.726	0.159
	(12.049)	(13.188)		
Female	0.434	0.465	-0.031	0.445
	(0.496)	(0.499)		
Full-time	0.761	0.703	0.058	0.087
	(0.427)	(0.458)		
Income	4.206	4.127	0.079	0.509
	(1.412)	(1.461)		
Political	-0.136	0.042	-0.178	0.225
	(1.889)	(1.940)		
Education:				
High school	0.083	0.062	0.021	0.309
	(0.276)	(0.242)		
Some college	0.130	0.150	-0.020	0.446
	(0.337)	(0.358)		
College	0.501	0.504	-0.003	1.000
	(0.501)	(0.501)		
Master's/PhD	0.286	0.283	-0.003	1.000
	(0.453)	(0.451)		
Observations	339	353		

Notes: This table describes the background characteristics of the participants from the manipulation check survey. They were randomly assigned to either reading the job instructions from the control condition or disinformation treatment. "Age" is a continuous variable that measures a subject's age. "Female" is a dummy variable that has a value of 1 if a subject is female. "Full-time" is a dummy variable that measures if a subject is employed full-time in their job. "Income" is a variable that ranges from 1 to 7, where a subject is asked to report its household income, compared to the average income in the US. "Political" is a variable that ranges from -3 to 3, and asks if a subject is politically liberal (-3) or conservative (3). "High school," "Some college," "College," and "Master's/PhD" are dummy variables for a subject's highest level of education. Due to very few observations with a Master's degree or higher, we combine "Master's" and "PhDs" into one category.

1.3 Downstream Consequences Survey Design Details

We conducted the downstream consequences survey on Prolific in April 2021 using Qualtrics. Prolific provided us with a representative sample of the US population with respect to age ("18-27," "28-37," "38-47," "48-57," "58 or older"), sex ("male," "female"), and ethnicity ("White," "Mixed," "Black," "Asian," "Other"). Respondents were paid $0.87 for completing the survey, which was expected to take 5 minutes or less to complete (median completion time was 4 minutes and 11 seconds). Using a between-subject design, respondents were randomly assigned to view a graph showing the official data or a graph created in the disinformation treatment. If a respondent successfully completed the attention check, they were asked to imagine viewing the graph on social media (e.g., Facebook or Twitter). Note, they were not told that the graph was created as part of a field experiment at this point (they were debriefed after the completion of the survey).

After viewing the graph, respondents were asked several questions to understand the downstream consequences of viewing such a graph on social media. In particular, the questions assessed the downstream consequences related to COVID-19 risk perceptions and behaviors. To assess people's behavioral intentions that researchers identified as increasing people's risk of catching and spreading COVID-19, we asked them on a 7-point scale from 1 ("Very uncomfortable") to 7 ("Very comfortable"): "After seeing this graph, would you feel comfortable or uncomfortable doing each of the following in California?" (1) Going out to the grocery store, (2) Eating out in a restaurant, (3) Attending an indoor sporting event or concert, (4) Visiting with a close friend or family member inside their home, (5) Supporting mandatory mask wearing on public places in California, (6) Travel to California for a pre-paid trip in the next month. Note, since the graphs in the field experiment used COVID-19 data from California, the questions pertained to the respondents' behavior if they were in that state.

To measure the impact of viewing the graph on COVID-19 risk perceptions, we asked respondents the following two questions on a 7-point scale from 1 ("Not at all worried") to 7 ("Extremely worried"): "After seeing this graph, how worried are you about the health consequences of Covid19 for you?" and "After seeing this graph, how worried are you that the Covid19 mutation will lead to a new wave of infections?" We further asked them about the likelihood of getting a COVID-19 vaccine on a 7-point scale from 1 ("Much less likely to be vaccinated") to 7 ("Much more likely to be vaccinated").

To understand the impact of viewing the COVID-19 graph on people's trust in media and public health officials, the following two questions were asked. For the first question, on a 7-point scale from 1 ("No trust at all") to 7 ("a lot of trust"), we asked: "After seeing this graph, how would you rate your trust in the mainstream media's reporting of Covid19?" For the second question, on a 7-point scale from 1 ("Poor") to 7 ("Excellent"), we asked, "After seeing this graph, how would you rate your trust in the job

Public health officials, such as those at the CDC (Centers for Disease Control and Prevention), are doing responding to the Covid19 outbreak?"

To assess people's willingness to share the graph on social media, we asked on a 7-point scale from 1 ("Extremely unlikely") to 7 ("Extremely likely"): "After seeing this graph, how likely would you be to share this graph with your friends and/or followers on social media?" We also asked questions about respondents' likelihood to fact-check the information in the graph, political orientation, media consumption, and vaccination status. The complete survey can be found in section 4.3.

For this survey, we pre-registered 800 participants (i.e., 400 per condition) based on a power calculation. According to our power calculation, we needed at least 394 participants in each condition to detect a small effect (Cohen's d = 0.2) for people's risk perceptions and behaviors with a power of 0.8 and significance threshold of 0.05 (two-sided). 997 respondents started the survey, 794 successfully passed the attention check and completed the survey. Thus, our final sample has 794 participants.

A balance check of the background characteristics shows that the variables are balanced across conditions. Additionally, none of the differences across conditions are significant at the 5% level (see Table A3).

Table A3: Balance Check for the Downstream Consequences Survey

	Control	Disinformation		
	(1)	(2)	(3)	(4)
	Mean (SD)		Δ	p-value
Age	47.201	47.426	-0.225	0.844
	(16.021)	(15.947)		
Female	0.522	0.534	-0.012	0.776
	(0.500)	(0.499)		
Full-time	0.244	0.277	-0.033	0.332
	(0.430)	(0.448)		
Political	-0.634	-0.643	0.009	0.854
	(1.762)	(1.803)		
White	0.720	0.728	-0.008	0.812
	(0.450)	(0.445)		
Observations	393	401		

Notes: This table describes the background characteristics of the participants from the downstream consequences survey. They were randomly assigned to either viewing a graph from the control condition

or disinformation treatment. “Age” is a continuous variable that measures a subject’s age. “Female” is a dummy variable that has a value of 1 if a subject is female. “Full-time” is a dummy variable that measures if a subject is employed full-time in their job. “Political” is a variable that ranges from -3 to 3, and asks if a subject is politically liberal (-3) or conservative (3). “White” is a dummy that indicates if a subject’s race is white.

1.4. Platform Intervention Survey Design Details

For the platform intervention survey, which was conducted in January 2022 using Qualtrics, we recruited MTurk workers to leverage their contextual experience and expertise as forecasters. The participants had the same qualifications as the workers from the field experiment and manipulation check (US residents, 500 or more jobs completed, approval rate of 95% or higher). They were paid $0.75 for completing the survey, which was expected to take 5 minutes or less (median completion time was 6 minutes and 42 seconds). If a respondent successfully completed the attention check, they were shown the instructions that MTurk workers in the disinformation treatment of the field experiment received. They were then asked to predict the number of workers (out of 100) who would accept and work on the disinformation job.

Next, to assess the effectiveness of interventions in deterring workers from accepting and completing the disinformation job, forecasters were shown five different platform interventions. Using a within-subjects design, we randomized the order of the interventions for each forecaster. For the “Reminder” intervention, we asked them to imagine: “... MTurk implemented a policy requiring all participants to reaffirm, before accepting each new HIT, that they agree to the current MTurk policy stating: ‘You may not use, or encourage others to use, MTurk for any illegal, harmful, fraudulent, infringing, or objectionable activities.’” We used this intervention because research suggests that providing reminders to people about expected behavior can be an effective way of eliciting compliance (Beshears and Kosowsky, 2020).

It is possible workers on MTurk may not know what is considered as unethical behavior. To combat this shortcoming, we asked forecasters to imagine: “... a policy change that requires all MTurkers to watch a short training video, showing examples of HITs violating MTurk’s terms of service, when registering for the platform and then once a year.” The “Training” intervention is motivated by studies suggesting that mandating workers to take a training session can make people aware that certain conduct, which they believe is appropriate, is unethical (Lindsey et al. 2015).

Studies show that establishing effective whistleblowing programs can encourage people to report unethical behavior (Dungan et al. 2015). Although it is possible to report suspicious jobs on MTurk, the platform does not provide any incentives for a worker to report an unethical job. Thus, workers currently have to be willing to sacrifice their time and pay when reporting such a job. Correspondingly, for the

"Whistleblowing" intervention, we asked forecasters to imagine: "... MTurk implemented a policy that paid workers 10% of the HIT's reward (for example, $0.10 for a $1.00 HIT) for clicking on the 'Report this HIT' when a HIT violates Amazon Mechanical Turk's Term of Service. Inaccurately reporting a HIT would not lead to a payment."

The "Peer Information" intervention was motivated by research demonstrating that people can change their behavior when they become aware of how their peers are behaving (Dimant 2019). Accordingly, the Peer Information condition asked forecasters to envision: "... for each HIT, Mturk also provides information of how many times the HIT has been viewed." We also indicated that "the follow-up HIT has a high view count, but low completion rate" to highlight that other workers are not taking the disinformation job.

Finally, recent studies demonstrate that increasing workers' accountability in online labor markets can be effective at deterring unethical behavior (Brink et al. 2019). As a result, the "Accountability" intervention asked forecasters to envision that MTurk workers are suspended from the platform when a task they worked on was flagged as unethical: "Now imagine MTurk expands this policy such that if workers complete a HIT that violates Amazon Mechanical Turk's Term of Service, both requesters and workers will be suspended."

Before eliciting forecasters' background information, we asked them to report how confident they are in their predictions on a 7-point scale. The complete survey can be found in section 4.4.

For this survey, we pre-registered 400 participants based on a power calculation. According to our power calculation for the within-subjects design, we needed about 367 subjects to detect a small effect (Cohen's $d = 0.2$) of an intervention's predicted effectiveness with a power of 0.8 and significance threshold of 0.05 (two-sided). For our power calculation, we assume a correlation of 0.07 between paired observations (i.e., predictions from the same individual). This number was obtained from a pilot study, which showed that the smallest correlation between the status quo prediction and the prediction of an intervention was 0.07. The pilot study further revealed a time-trend in the predictions. As a result, to account for this, we increased the target sample size to 400. 578 forecasters started the survey, 431 successfully passed the attention check, and 400 both completed the survey and successfully submitted it. Table A4 below presents descriptive statistics of our sample.

Table A4: Descriptive Statistics for the Platform Intervention Survey

	Mean
	(SD)
Age	39.545
	(10.804)
Female	0.428
	(0.495)
Full-time	0.775
	(0.418)
Income	4.365
	(1.285)
Political	0.230
	(1.779)
Education:	
High school	0.068
	(0.251)
Some College	0.110
	(0.313)
College	0.523
	(0.500)
Master's/PhD	0.300
	(0.458)
Observations	400

Note: This table describes the background characteristics of the participants from the platform intervention survey. "Age" is a continuous variable that measures a subject's age. "Female" is a dummy variable that has a value of 1 if a subject is female. "Full-time" is a dummy variable that measures if a subject is employed full-time in their job. "Income" is a variable that ranges from 1 to 7, where a subject is asked to report its household income, compared to the average income in the US. "Political" is a variable that ranges from -3 to 3, and asks if a subject is politically liberal (-3) or conservative (3). "High school," "Some college," "College," and "Master's/PhD" are dummy variables for a subject's highest level of education. Due to very few observations with a Master's degree or higher, we combine "Master's" and "PhDs" into one category.

2. IRB Approval and Assessment of Field Experiment's Ethical Risks

Our project was approved by both University of Michigan's IRB (HUM00179761) and Erasmus University's IRB (IRB-E Approval 2019-06). While only the University of Michigan's IRB approval was required because the research team from the University of Michigan was responsible for collecting data for this project, given the potential ethical risks of the field experiment, we also sought and received IRB approval from Erasmus' IRB to ensure our project was vetted by multiple experts from two countries. Below we detail how we assessed and took efforts to minimize the study's ethical risks, focusing on our field experiment.

To measure the causal effect of a job's ethicality on a person's willingness to produce disinformation, our field experiment used incomplete disclosure and delayed debriefing. In accordance with IRB's standards for protecting human subjects, incomplete disclosure and delayed debriefing is allowed when: (1) the costs are minimal, (2) the subjects are not exposed to emotional or physical pain, (3) the research cannot be performed in another way, and (4) the benefits are significant. Below, we include pertinent information we considered for each factor when we designed our field experiment, particularly as it relates to the disinformation treatment:

(1) **Minimal costs:** Workers in each treatment were compensated fairly for their work. Recent work suggests that the average employer pays over $11 per hour of work on MTurk (Hara et al. 2018). Further, scholars suggest that in determining fair, acceptable wages it is important to observe the payment rate for similar task requests to ensure the payment is not too high or low (Lovett et al. 2018). Based on pilot studies, we estimated that it would take workers about six minutes to complete the job. Accordingly, we set the payment rate at $1.20, which equates to paying a worker $12 per hour to complete the job. In the lower-pay condition, workers were offered half the wage ($0.60) they were offered for the first job. Although workers in the lower-pay treatment were offered considerably less than what they were paid for the first job, they still earned more than the median hourly earnings workers receive on MTurk (Toxtli et al. 2021).

Further, workers could simply ignore the job posting (first job) and job offer invitation (second job), respectively. Thus, if workers did not want to work on these jobs, they could just move on to other jobs posted on MTurk without losing noticeable time.

(2) **Not Exposed to Emotional or Physical Pain:** Amazon Mechanical Turk is configured such that working on any job is completely voluntary: our subjects could have skipped the disinformation job or withdrawn from it at any time, without any repercussions we are aware of. Further, to minimize the potential of any unintended consequences, we did not post the graphs on social media and debriefed participants to ensure they knew that their graphs were created only for research purposes. Additionally, we considered three scenarios in which the subject could have voluntarily withdrawn from the study:

a) Before Accepting the Job: Every worker was told about the unethical nature of the disinformation job before they accepted working on it. For example, in our job invitation, workers in the disinformation treatment were sent an email, using MTurk's application program interface (API), containing the following text:

> *You recently completed a HIT visualizing Covid19 infections. I'm inviting you for a follow-up HIT to create a similar graph for Covid19 deaths, but this time I need you to reduce the numbers compared to what the "original" numbers would look like. The HIT will pay $1.20 and it will again take about 6 minutes to complete.*
>
> *When you are done, the Covid19 death curve must look much flatter than the original entries. I don't care how you change the numbers as long as someone can tell that the Covid19 death rate looks less worrying after viewing the manipulated graph.*
> *I will publicly post this new graph on Facebook and Twitter.*
>
> *Please search for "[redacted]" or "Simple Data Visualization (Follow-Up)" to accept the HIT.*

Additionally, on MTurk, workers are only evaluated after they submit their work on a job, not on which jobs they choose or do not choose to accept. As a result, workers' participation was voluntary, they were not compelled by the employer to complete the job, and could delete, ignore, or skip the email or job, without any discernible repercussions.

Before accepting the disinformation job offer, three out of 400 workers emailed us asking about our intentions for offering this job. Anticipating that this could occur, before the field experiment, we prepared and subsequently sent the following email:

> *This is a study manipulation and part of an academic study that has been approved by an ethics board. Because we are hoping to get respondents' spontaneous responses, we are planning to debrief participants after the study is complete. In the meantime, please do not share any information about this study with others to ensure other respondents are not influenced. Please do not hesitate to reach out to me if you have any questions or concerns.*

b) After Accepting the Job: Even after starting a job, workers were free to withdraw for any reason (for example, if they changed their minds about how comfortable they felt working on the disinformation job). Workers were able to withdraw without explicitly communicating with us and could simply cancel the job through MTurk. If a worker took this step, the action would not show up on a worker's profile and we did not receive any notification or indication of this action. We were simply allowed to see which workers completed each job and were not given any information about workers who did not complete a job. Note, if a worker did not complete the job after starting it, MTurk does not allow employers to pay money for the job. As a result, workers do not earn pay for partial work they completed, but decided not to submit.

c) After Project Completion: The third situation is if workers later regretted working on the disinformation job and contacted us about the job. This situation did not occur during the project or after debriefing workers. If this situation had occurred during the study, our protocol was to debrief the participant immediately and give them the option to remove their data from the study.

(3) **Research Cannot be Performed Another Way:** A limitation of the existing literature on immoral work is that subjects typically know that they are being observed. Accordingly, they might behave differently (e.g., more ethically) than in real life situations, limiting our understanding about this important topic. In our field experiment, we believe that requesting explicit consent before data collection was completed would have significantly changed participants' behavior, hence the need to delay debriefing until after our field experiment was completed.

(4) **Benefits are Significant:** As discussed in our paper, misinformation and disinformation has emerged as one of the leading factors impacting people's beliefs and actions, especially in the age of social media. Not only is this evident with COVID-19, but also with climate change, vaccination, politics, immigration, and essentially every issue in public discourse. Understanding the production of disinformation is an important, yet overlooked, aspect in understanding and ultimately addressing this phenomenon, particularly informing our approach to discouraging the production of disinformation.

Taken together, in consultation with two different University IRB's, we took concerted efforts to assess and design our field experiment in a way that minimizes the project's risks to subjects.

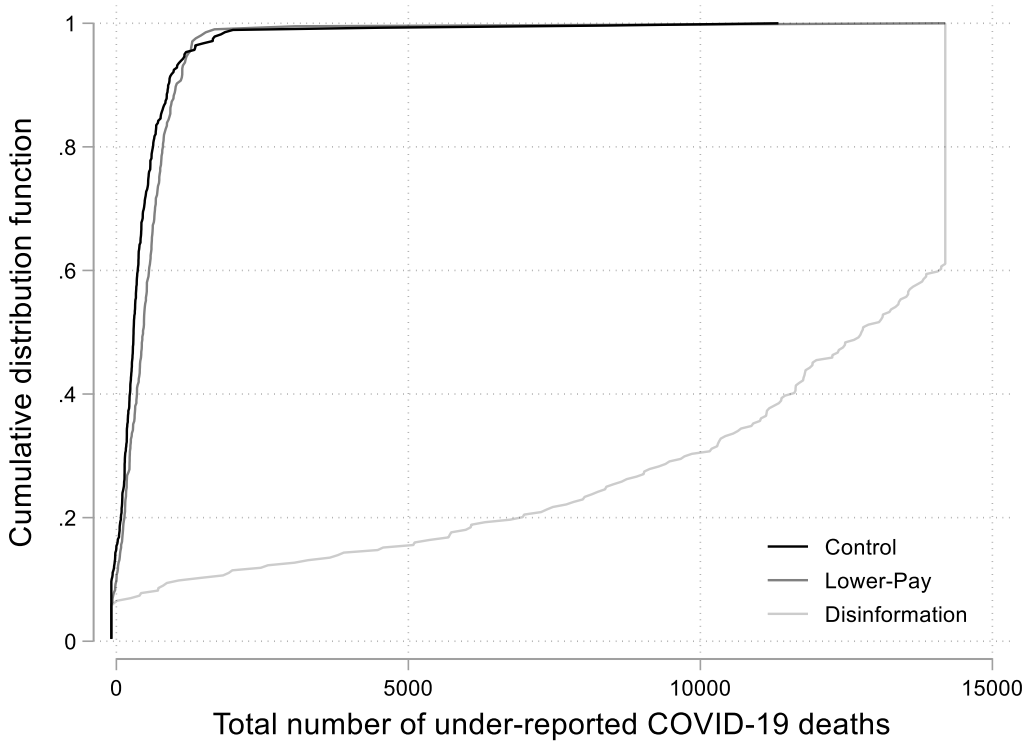
3. Analysis

In this section, we present additional analyses for each study providing further insight into the results presented in the main text. Unless noted, we report the results from OLS regressions. For our main treatment of interest, the disinformation treatment, we also report p-values adjusted for multiple hypothesis testing (MHT) using the Bonferroni correction method. Note, the Bonferroni correction is a conservative approach to correct for MHT (List et al. 2019).

3.1 Field experiment

Figure A1 displays the cumulative distribution function of the total number of under-reported COVID-19 deaths for each condition.

Figure A1: Under-reported COVID-19 deaths (cumulative distribution)



Notes: This figure shows the cumulative distribution function of the total number of under-reported COVID-19 deaths by treatment. The data is winsorized to the 5th and 95th percentile.

Table A5 presents the regression results of our main outcomes from the field experiment. Column (1) shows that, while in the control condition about 70% of the workers accepted and completed the second job, the job acceptance rate decreased by 9 percentage points in the disinformation treatment ($P = 0.007$, t-test) and 19 percentage points in the lower-pay treatment ($P < 0.001$, t-test), respectively. Column (2) shows that the results do not change meaningfully when we use a Probit instead of OLS regression due to the binary outcome variable.

Table A5: Job Acceptance and Under-reported Deaths in the Field Experiment

	(1)	(2)	(3)	(4)	(5)
Dep. variable	Job acceptance (=1)		% Deaths reported		
Disinformation	-0.091***	-0.093***	-50.523***	-54.216***	-49.059***
	(0.034)	(0.034)	(1.937)	(1.833)	(1.589)
Lower-pay	-0.187***	-0.186***	3.315	3.400	-0.564

	(0.034)	(0.033)	(4.450)	(4.439)	(0.465)
Constant	0.701***		98.559***	98.909***	98.028***
	(0.023)		(0.506)	(0.299)	(0.247)
Observations	1,197	1,197	728	728	728
P-value corrected for MHT (disinformation treatment)	0.014	0.012	<0.001	<0.001	<0.001

Notes: OLS regressions in columns 1 and 3-5. Probit regression (marginal effects) in column 2. Robust standard errors in parenthesis. Columns 1 and 2 report the share of workers who accepted the second job by treatment. Columns 3-5 show how many COVID-19 deaths were reported (in percent) in each condition (i) for all 12 weeks (column 3), (ii) after week six (which is the time period workers were instructed to focus on) (column 4), (iii) winsorizing the data at the 5th and 95th percentile for all 12 weeks (column 5). Significance levels: * p < 0.10, ** p < 0.05, *** p < 0.01. For the disinformation treatment, we additionally report p-values corrected for multiple hypothesis testing (MHT) using the Bonferroni correction method (five hypotheses).

Column (1) in Table A6 shows that it took workers about 4.5 hours longer to accept the disinformation job compared to the control job (P = 0.126, t-test), suggesting that at least some workers hesitated before accepting the disinformation job. In fact, workers waited about the same amount of time before accepting the job as those in the lower-pay treatment (P = 0.780, t-test). Column (2) in this table shows that the results remain qualitatively the same when we winsorize the data at the 5th and 95th percentile to account for outliers.

Table A6: Duration from Job Invitation to Acceptance

	(1)	(2)
Dep. variable	Time to job acceptance (hours)	
Disinformation	4.542	2.939
	(2.965)	(2.219)
Lower-pay	5.491*	2.641
	(3.196)	(2.278)
Constant	57.771***	54.226***
	(1.927)	(1.478)

Observations	728	728
P-value corrected for MHT (Disinformation treatment)	0.252	0.372

Notes: OLS regressions with robust standard errors in parenthesis. This table shows how much time elapsed (in hours) before a worker accepted the second job by treatment. Column (1) reports the raw data, and column (2) uses winsorized data at the 5th and 95th percentile. Significance levels: * p < 0.10, ** p < 0.05, *** p < 0.01. For the disinformation treatment, we additionally report p-values corrected for multiple hypothesis testing (MHT) using the Bonferroni correction method (two hypotheses).

Returning to Table A5, column (3) shows that workers in the disinformation treatment reduced the number of deaths by more than 50% ($P < 0.001$, t-test). This result suggests that workers did not just superficially complete the job, but made a concerted effort to make the COVID-19 death rate look less worrying. This is also reflected when analyzing the amount of time workers took to complete the job. Column (1) in Table A7 shows that workers in the disinformation treatment did not complete the job significantly faster than the control group ($P = 0.256$, t-test), regardless of whether we winsorize the completion time at the 5th and 95th percentile ($P = 0.264$, t-test).

Although workers did not receive detailed instructions on how much to reduce the number of deaths, they were told to especially "reduce" the number of deaths after a specific date so that the curve would look flatter compared to a graph using official data. In the official data, the specific date corresponds with a sharp increase in the number of deaths. Focusing on the numbers after this specific date, we observe in Table A5's column (4) that workers in the disinformation treatment under-reported the number of deaths by 54.2% compared to the control condition ($P < 0.001$, t-test). Column (5) further shows that the results from column (3) are robust to winsorizing the reported death rates at the 5th and 95th percentile. This suggests that the results are not just driven by outliers (e.g., three percent of workers graphed zero deaths for each week).

Focusing on the lower-pay treatment, column (3) in Table A5 shows that workers reported similar death numbers as those in the control condition ($P = 0.457$, t-test). Although they reported similar numbers, workers in the lower-pay treatment completed the job significantly faster compared to the control group, as shown in column (1) in Table A7 ($P = 0.001$, t-test).

Table A7: Duration for Completing the Job

	(1)	(2)

Dep. variable	Work time (seconds)	
Disinformation	-12.017	-9.974
	(10.568)	(10.052)
Lower-pay	-33.527***	-30.763***
	(10.453)	(10.021)
Constant	235.595***	234.670***
	(7.098)	(6.796)
Observations	728	728
P-value corrected for MHT (Disinformation treatment)	0.512	0.642

Notes: OLS regressions with robust standard errors in parenthesis. This table shows how much time (in seconds) it took workers to complete the second job by treatment. Column (1) reports the raw data, and column (2) uses winsorized data at the 5th and 95th percentile. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For the disinformation treatment, we additionally report p-values corrected for multiple hypothesis testing (MHT) using the Bonferroni correction method (two hypotheses).

3.2 Manipulation Check

Table A8 reports p-values from nonparametric tests, corrected for multiple hypothesis testing (MHT, five hypotheses) using the Bonferroni correction method. None of the corrected p-values crosses the 0.001 threshold.

Table A8: P-values from Nonparametric Tests Corrected for MHT (Manipulation Check Survey)

	P-value (uncorrected)	P-value corrected for MHT (Bonferroni)
Immoral	< 0.001	< 0.001
Job acceptance	< 0.001	< 0.001
Predicted job acceptance	< 0.001	< 0.001
Lowest Pay	< 0.001	< 0.001
Perceived responsibility	< 0.001	< 0.001

Notes: "Immoral" is a dummy variable that takes the value of 1 if a subject perceives the job as unethical, and 0 otherwise. "Accept job" is a dummy variable that takes the value of 1 if a subject indicates they would complete the job. "Prediction" is a subject's prediction of how many workers (out of 100) would take the job. "Lowest Pay" is the lowest wage they would accept to work on the job using a sliding scale between $0 and $5. "Responsible" indicates how personally responsible a subject would feel about how the graph created from the job could affect people who see it on social media, from 1 ("Not at all responsible") to 7 ("Fully responsible").

As pre-registered, Table A9 reports the regression results for all five outcomes variables. Compared to the non-parametric analysis in the main text, the regressions allow us to control for participants' background characteristics.

Table A9: Perceived Ethicality of the Job

Dep. variable	(1) Immoral	(2) Job acceptance	(3) Predicted job acceptance	(4) Lowest Pay	(5) Perceived Responsibility
Disinformation	0.381***	-0.247***	-8.745***	0.600***	0.782***
	(0.029)	(0.029)	(1.814)	(0.110)	(0.137)
Age	0.001	-0.004***	-0.054	-0.007	-0.004
	(0.001)	(0.001)	(0.073)	(0.005)	(0.006)
Female	0.017	-0.072**	-2.144	-0.092	0.259*
	(0.030)	(0.030)	(1.859)	(0.113)	(0.141)
Full-time	-0.128***	0.163***	1.071	0.155	-0.083
	(0.039)	(0.042)	(2.415)	(0.146)	(0.186)
Income	0.005	0.002	0.507	0.196***	0.415***
	(0.012)	(0.013)	(0.769)	(0.050)	(0.060)
Political	-0.031***	0.033***	0.743	0.004	-0.083**
	(0.008)	(0.008)	(0.484)	(0.032)	(0.040)
Education:					
Some college	0.037	-0.123	-9.061*	0.020	0.090
	(0.069)	(0.082)	(4.630)	(0.251)	(0.372)
College degree	-0.027	0.056	-5.193	-0.214	0.154
	(0.063)	(0.074)	(4.129)	(0.220)	(0.336)
Master's/PhD	-0.056	0.052	-5.390	-0.012	0.480
	(0.067)	(0.078)	(4.428)	(0.239)	(0.354)
Constant	0.090	0.924***	80.566***	1.308***	2.316***
	(0.091)	(0.102)	(5.494)	(0.348)	(0.462)
Observations	690	690	690	690	690
P-value corrected for MHT (Disinformation treatment)	<0.001	<0.001	<0.001	<0.001	<0.001

Notes: OLS regressions with robust standard errors in parenthesis. This table shows people's perceived ethicality of the second job by treatment. The dependent variable "Immoral" is a dummy variable that takes the value of 1 if a subject perceives the job as unethical, and 0 otherwise. "Job Acceptance" is a dummy

variable that takes the value of 1 if a subject indicates they would complete the job. "Predicted Job Acceptance" is a subject's prediction of how many workers (out of 100) would take the job. "Lowest Pay" is the lowest wage they would accept to work on the job using a sliding scale between $0 and $5. "Perceived Responsibility" indicates how personally responsible a subject would feel about how the graph created from the job could affect people who see it on social media, from 1 ("Not at all responsible") to 7 ("Fully responsible"). "Disinformation" is a dummy variable that has the value of 1 if a subject was shown the instructions for the disinformation job, and 0 otherwise. "Age" is a continuous variable that measures a subject's age. "Female" is a dummy variable that has the value of 1 if a subject is female. "Full-time" is a dummy variable that measures if a subject is employed full-time in their job. "Income" is a variable that ranges from 1 to 7, where a subject is asked to report its household income, compared to the average income in the US. "Political" is a variable that ranges from -3 to 3, and asks if a subject is politically liberal (-3) or conservative (3). "High school," "Some college," "College," and "Master's/PhD" are dummy variables for a subject's highest level of education. Due to very few observations with a Master's degree or higher, we combine "Master's" and "PhDs" into one category. When we control for background characteristics, the sample size decreases by two because two subjects did not report their age. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For the disinformation treatment, we additionally report p-values corrected for multiple hypothesis testing (MHT) using the Bonferroni correction method (five hypotheses).

Column 1 focuses on the moral acceptability to work on the job. Recall that, as pre-registered, we combined the answer options "morally acceptable" and "not a moral issue" to create a binary variable identifying the share of workers who found the job unethical in each condition. The disinformation treatment increases the share of respondents who view the job as morally unacceptable by 38 percentage points ($P < 0.001$, t-test, adjusted for MHT), from 6% in the control condition to 45% in the disinformation treatment.

Column 2 reports respondents' willingness to accept and complete the job. The disinformation job decreases respondents' willingness by 25 percentage points ($P < 0.001$, t-test, adjusted for MHT), from 89% in the control to 63% in the disinformation treatment. Note, we combined the answer options "no" and "I am not sure" (overall, only 8% of the participants responded with "I am not sure"). This allows us to estimate the share of workers who are willing to do the job and thus we can compare this response to the field experiment's results. The results indicate that the respondents are well calibrated (recall, in the field experiment 70% completed the control job and 61% completed the disinformation job).

Column 3 reports respondents' predictions about how many other workers would accept and work on the job. The respondents predicted that out of 100 workers, 9 fewer workers would accept and complete the disinformation job ($P < 0.001$, t-test, adjusted for MHT). This corresponds to a decrease from 75 workers in the control condition to 66 workers in the disinformation treatment. Thus, the results are qualitatively similar, regardless of whether we ask participants about their own willingness to do the job or to predict what others would do.

Column 4 reports respondents' preference for the lowest pay they would be willing to receive to complete the job. The disinformation job increases the minimum acceptable pay by $0.60 ($P < 0.001$, t-test, adjusted for MHT). This corresponds to workers requiring a 34% wage premium for the disinformation job.

Finally, column 5 reports the extent to which respondents would feel personally responsible for how the graph affects others, if they worked on the job. The disinformation job increases respondents' feeling of responsibility by 0.8 points on a 7-point scale ($P < 0.001$, t-test, adjusted for MHT). This result suggests that respondents understood the social harm that the graph created in the disinformation job could have on others, if it was shared more widely on social media.

Even though respondents were anonymous, given the unethical nature of the disinformation job, answers to the five questions may be biased by respondents' social image concerns. As a result, we conducted a post-hoc robustness check restricting the analysis to those who are less likely to provide socially desirable answers (i.e., those with a below median IM score). Shown in Table A10, the results remain qualitatively the same with the restricted sample.

Table A10: Perceived Ethicality of the Job (Below-Median IM Score)

Dep. variable	(1) Immoral	(2) Job acceptance	(3) Predicted job acceptance	(4) Lowest Pay	(5) Perceived Responsibility
Disinformation	0.290***	-0.183***	-5.433**	0.410***	0.689***
	(0.043)	(0.038)	(2.398)	(0.149)	(0.186)
Age	0.001	-0.002	0.048	-0.002	-0.020**
	(0.002)	(0.002)	(0.106)	(0.007)	(0.009)
Female	0.008	-0.086**	-1.639	0.108	0.371*
	(0.044)	(0.038)	(2.523)	(0.149)	(0.193)
Full-time	-0.142**	0.201***	-2.082	0.223	0.107
	(0.061)	(0.065)	(3.430)	(0.216)	(0.288)
Income	-0.007	0.025	0.474	0.150**	0.360***
	(0.017)	(0.017)	(0.975)	(0.065)	(0.080)
Political	-0.021*	0.032***	0.497	0.004	-0.055
	(0.011)	(0.011)	(0.688)	(0.047)	(0.059)
Education:					
Some college	-0.065	-0.075	-10.723	0.448	0.304
	(0.123)	(0.138)	(6.534)	(0.340)	(0.540)
College degree	-0.131	0.071	-11.574*	0.103	0.476
	(0.109)	(0.120)	(5.903)	(0.268)	(0.489)
Master's/PhD	-0.148	0.079	-10.126	0.077	0.849*
	(0.116)	(0.123)	(6.338)	(0.299)	(0.513)
Constant	0.318**	0.709***	82.359***	0.883**	2.511***
	(0.148)	(0.153)	(8.164)	(0.433)	(0.611)
Observations	335	335	335	335	335
P-value corrected for MHT (Disinformation treatment)	<0.001	<0.001	0.120	0.030	<0.001

Notes: OLS regressions with robust standard errors in parenthesis. This table shows people's perceived ethicality of the second job by treatment for those who are less likely to provide socially desirable answers (i.e., those with a below median IM score). The dependent variable "Immoral" is a dummy variable that takes the value of 1 if a subject perceives the job as unethical, and 0 otherwise. "Job Acceptance" is a

dummy variable that takes the value of 1 if a subject indicates they would complete the job. "Predicted Job Acceptance" is a subject's prediction of how many workers (out of 100) would take the job. "Lowest Pay" is the lowest wage they would accept to work on the job using a sliding scale between $0 and $5. "Perceived Responsibility" indicates how personally responsible a subject would feel about how the graph created from the job could affect people who see it on social media, from 1 ("Not at all responsible") to 7 ("Fully responsible"). "Disinformation" is a dummy variable that has the value of 1 if a subject was shown the instructions for the disinformation job, and 0 otherwise. "Age" is a continuous variable that measures a subject's age. "Female" is a dummy variable that has the value of 1 if a subject is female. "Full-time" is a dummy variable that measures if a subject is employed full-time in their job. "Income" is a variable that ranges from 1 to 7, where a subject is asked to report its household income, compared to the average income in the US. "Political" is a variable that ranges from -3 to 3, and asks if a subject is politically liberal (-3) or conservative (3). "High school," "Some college," "College," and "Master's/PhD" are dummy variables for a subject's highest level of education. Due to very few observations with a Master's degree or higher, we combine "Master's" and "PhDs" into one category. When we control for background characteristics, the sample size decreases by two because two subjects did not report their age. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For the disinformation treatment, we additionally report p-values corrected for multiple hypothesis testing (MHT) using the Bonferroni correction method (five hypotheses).

Given that political orientation shapes how people have understood and responded to the COVID-19 pandemic (Grossman et al. 2020), we analyzed whether respondents' partisanship affected their perceived ethicality of the disinformation job. Indeed, more conservative respondents considered the disinformation job as less unethical; as shown in Table A11, column 1 shows that for each additional point on the political orientation scale (with higher values indicating being more conservative), the likelihood of perceiving the disinformation job as morally unacceptable decreases by 7 percentage points ($P < 0.001$, t-test). In line with this finding, more conservative respondents also indicated they would be more likely to accept and work on the disinformation job, require lower compensation to do so, feel less personally responsible for how the graph affects others, and expect more people to accept working on the disinformation job (all $P < 0.001$, t-tests, see columns 2 to 5). Overall, the results suggest that partisan bias in how people perceived the COVID-19 pandemic also matters for our study.

Table A11: Perceived Ethicality of the Job (by Political Orientation)

Dep. variable	(1) Immoral	(2) Job acceptance	(3) Predicted job acceptance	(4) Lowest Pay	(5) Perceived Responsibility
Disinformation	0.377***	-0.244***	-8.647***	0.590***	0.773***
	(0.029)	(0.029)	(1.809)	(0.109)	(0.137)
Disinf. × Political	-0.065***	0.064***	2.037**	-0.223***	-0.184**
	(0.014)	(0.014)	(0.880)	(0.060)	(0.073)
Age	0.001	-0.004***	-0.053	-0.007	-0.005
	(0.001)	(0.001)	(0.072)	(0.005)	(0.006)
Female	0.021	-0.076**	-2.280	-0.077	0.272*
	(0.030)	(0.030)	(1.848)	(0.112)	(0.141)
Full-time	-0.122***	0.157***	0.896	0.174	-0.067
	(0.039)	(0.041)	(2.398)	(0.143)	(0.185)
Income	0.003	0.004	0.570	0.189***	0.409***
	(0.012)	(0.013)	(0.770)	(0.049)	(0.059)
Political	0.003	-0.000	-0.330	0.121***	0.014
	(0.007)	(0.008)	(0.590)	(0.039)	(0.057)
Education:					
Some college	0.041	-0.127	-9.180**	0.033	0.101
	(0.070)	(0.083)	(4.643)	(0.251)	(0.370)
College degree	-0.019	0.048	-5.436	-0.187	0.176
	(0.064)	(0.076)	(4.177)	(0.219)	(0.336)
Master's/PhD	-0.048	0.045	-5.641	0.015	0.503
	(0.068)	(0.079)	(4.473)	(0.238)	(0.353)
Constant	0.092	0.922***	80.501***	1.315***	2.322***
	(0.091)	(0.103)	(5.523)	(0.352)	(0.464)
Observations	690	690	690	690	690
P-value corrected for MHT (Disinf.×Political)	<0.001	<0.001	0.105	<0.001	0.060

Notes: OLS regressions with robust standard errors in parenthesis. This table shows people's perceived ethicality of the second job by their political orientation. The dependent variable "Immoral" is a dummy variable that takes the value of 1 if a subject perceives the job as unethical, and 0 otherwise. "Job Acceptance" is a dummy variable that takes the value of 1 if a subject indicates they would complete the job. "Predicted Job Acceptance" is a subject's prediction of how many workers (out of 100) would take the job. "Lowest Pay" is the lowest wage they would accept to work on the job using a sliding scale between $0 and $5. "Perceived Responsibility" indicates how personally responsible a subject would feel about how the graph created from the job could affect people who see it on social media, from 1 ("Not at all responsible") to 7 ("Fully responsible"). "Disinformation" is a dummy variable that has the value of 1 if a subject was shown the instructions for the disinformation job, and 0 otherwise. "Age" is a continuous variable that measures a subject's age. "Female" is a dummy variable that has the value of 1 if a subject is female. "Full-time" is a dummy variable that measures if a subject is employed full-time in their job. "Income" is a variable that ranges from 1 to 7, where a subject is asked to report its household income, compared to the average income in the US. "Political" is a variable that ranges from -3 to 3, and asks if a subject is politically liberal (-3) or conservative (3). "Political x Disinformation" is the interaction term between political orientation and the disinformation job, indicating whether more conservative people have different views about the job's ethicality. "High school," "Some college," "College," and "Master's/PhD" are dummy variables for a subject's highest level of education. Due to very few observations with a Master's degree or higher, we combine "Master's" and "PhDs" into one category. When we control for background characteristics, the sample size decreases by two because two subjects did not report their age. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For the disinformation treatment, we additionally report p-values corrected for multiple hypothesis testing (MHT) using the Bonferroni correction method (five hypotheses).

3.3 Downstream Consequences

As pre-registered, Tables A12, A13, and A14 report the regression results for our main outcome variables from the downstream consequences survey. The regressions correspond to the results presented in Figure 3 of the main text. Compared to the non-parametric analysis, the regressions allow us to control for participants' background characteristics. We standardize each outcome variable to have a mean of 0 and a standard deviation of 1. This allows us to compare effect sizes across outcomes. At the bottom of the table, we report p-values adjusted for MHT using the Bonferroni correction method. We make this adjustment within each group of outcome variables, which are presented in separate tables (risky behaviors, risk perceptions, trust in institutions and media). For convenience, in Table A14, we show the results for "trust in institutions and media" and "sharing" in the same table. Results for "trust in institutions and media" are corrected for MHT (two hypotheses), but this correction is not needed for "sharing" (one hypothesis). For completeness, Table A15 provides the corresponding p-values from non-parametric tests.

Table A12: Effect of Manipulated COVID-19 Graph on Risky Behaviors

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Variable	Grocery	Restaurant	Concert	Visit a friend	Wear a mask	Travel
Disinformation	0.595***	0.442***	0.316***	0.376***	-0.060	0.501***
	(0.067)	(0.065)	(0.066)	(0.066)	(0.065)	(0.067)
Age	-0.008***	-0.013***	-0.010***	-0.012***	0.006***	-0.012***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Female	-0.003	-0.073	-0.074	-0.136**	0.125*	-0.082
	(0.068)	(0.067)	(0.068)	(0.067)	(0.066)	(0.068)
Political	0.114***	0.181***	0.190***	0.162***	-0.220***	0.106***
	(0.020)	(0.020)	(0.021)	(0.020)	(0.021)	(0.021)
White	0.215***	0.219***	0.277***	0.306***	-0.285***	0.248***
	(0.077)	(0.072)	(0.069)	(0.075)	(0.068)	(0.074)
Constant	-0.012	0.356***	0.276**	0.344***	-0.262**	0.260**
	(0.124)	(0.115)	(0.114)	(0.121)	(0.124)	(0.121)
Observations	775	775	775	775	775	775
P-value corrected for MHT (Disinformation treatment)	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Notes: OLS regressions with robust standard errors in parenthesis. All dependent variables were measured on a 7-point scale from 1 ("Very uncomfortable") to 7 ("Very comfortable"), and then standardized with mean 0 and standard deviation of 1. Disinformation is a dummy variable that has the value of 1 when subjects were shown the manipulated graph. "Age" is a continuous variable that measures a subject's age. "Female" is a dummy variable that has a value of 1 if a subject is female. "Political" is a variable that ranges from -3 to 3, and asks if a subject is politically liberal (-3) or conservative (3). "White" is a dummy that indicates if a subject's race is white. When we control for background characteristics, the sample size decreases by nineteen because they did not report their age or ethnicity (or both). Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For the disinformation treatment, we additionally report p-values corrected for multiple hypothesis testing (MHT) using the Bonferroni correction method (six hypotheses).

Table A13: Effect of Manipulated COVID-19 Graph on Risk Perceptions

	(1)	(2)	(3)
Dep. Variable	Health concerns	New variant	Vaccine intent
Disinformation	-0.240***	-0.312***	-0.447***
	(0.069)	(0.067)	(0.069)
Age	0.004*	0.008***	0.005**
	(0.002)	(0.002)	(0.002)
Female	0.005	0.092	-0.119*
	(0.071)	(0.068)	(0.069)
Political	-0.121***	-0.182***	-0.088***
	(0.021)	(0.020)	(0.021)
White	-0.310***	-0.274***	-0.235***
	(0.083)	(0.079)	(0.084)
Constant	0.079	-0.210*	0.174
	(0.131)	(0.123)	(0.131)
Observations	775	775	772
P-value corrected for MHT (Disinformation treatment)	0.003	<0.001	<0.001

Notes: OLS regressions with robust standard errors in parenthesis. The dependent variables "Health concerns" and "New variant" were measured on a 7-point scale from 1 ("Not at all worried") to 7 ("Extremely worried"). The dependent variable was measured on a 7-point scale from 1 ("Much less likely to be vaccinated") to 7 ("Much more likely to be vaccinated"). All dependent variables were standardized with mean 0 and standard deviation of 1. Disinformation is a dummy variable that has the value of 1 when subjects were shown the manipulated graph. "Age" is a continuous variable that measures a subject's age. "Female" is a dummy variable that has a value of 1 if a subject is female. "Political" is a variable that ranges from -3 to 3, and asks if a subject is politically liberal (-3) or conservative (3). "White" is a dummy that indicates if a subject's race is white. When we control for background characteristics, the sample size in columns 1 and 2 decreases by nineteen because they did not report their age or ethnicity (or both). The sample size decreases by an additional three subjects in column 3 due to nonresponse to the vaccination question. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For the disinformation treatment, we additionally report p-values corrected for multiple hypothesis testing (MHT) using the Bonferroni correction method (three hypotheses).

Table A14: Effect of Manipulated COVID-19 Graph on Trust in Institutions and Media, and Sharing Intent

Dep. Variable	(1) Trust in media	(2) Trust in health officials	(3) Sharing Intent
Disinformation	-0.174***	0.052	-0.137*
	(0.063)	(0.066)	(0.070)
Age	0.008***	0.004*	-0.000
	(0.002)	(0.002)	(0.002)
Female	-0.152**	-0.019	-0.124*
	(0.063)	(0.066)	(0.071)
Political	-0.268***	-0.226***	-0.010
	(0.019)	(0.021)	(0.022)
White	-0.238***	-0.175**	-0.417***
	(0.074)	(0.077)	(0.088)
Constant	-0.201*	-0.233*	0.442***
	(0.119)	(0.124)	(0.134)
Observations	775	775	775
P-value corrected for MHT (Disinformation treatment)	0.012	1.000	0.052

Notes: OLS regressions with robust standard errors in parenthesis. The dependent variable "Trust in media" was measured on a 7-point scale from 1 ("No trust at all" to 7 ("A lot of trust"). The dependent variable "Trust in health officials" was measured on a 7-point scale from 1 ("Poor" to "Excellent"). The dependent variable "Sharing intent" was measured on a 7-point scale from 1 ("Extremely unlikely" to 7 ("Extremely likely"). All dependent variables were standardized with mean 0 and standard deviation of 1. Disinformation is a dummy variable that has the value of 1 when subjects were shown the manipulated graph. "Age" is a continuous variable that measures a subject's age. "Female" is a dummy variable that has a value of 1 if a subject is female. "Political" is a variable that ranges from -3 to 3, and asks if a subject is politically liberal (-3) or conservative (3). "White" is a dummy that indicates if a subject's race is white. When we control for background characteristics, the sample size decreases by nineteen because they did not report their age or ethnicity (or both). Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For the disinformation treatment, we additionally report p-values corrected for multiple hypothesis testing (MHT) using the Bonferroni correction method (two hypotheses and one hypothesis).

Table A15 reports the corresponding p-values from nonparametric tests, corrected for MHT using the Bonferroni correction method. For convenience, we present the results from all four outcome variable groups in a single table.

Table A15: P-values from Nonparametric Tests Corrected for MHT (Downstream Consequences Survey)

Variable	P-value (uncorrected)	P-value corrected for MHT (Bonferroni)
Grocery	<0.001	<0.001
Restaurant	<0.001	<0.001
Concert	<0.001	<0.001
Visit friend	<0.001	<0.001
Wear mask	0.029	0.174
Travel	<0.001	<0.001
Health concerns	0.001	0.003
New variant	<0.001	<0.001
Vaccine intent	<0.001	<0.001
Trust in the media	0.003	0.006
Trust in health officials	0.560	1.000
Sharing intent	0.071	0.071

Notes: The first column shows uncorrected p-values and the second column shows p-values corrected for multiple hypothesis testing using the Bonferroni method (six, three, two, and one hypothesis).

Research suggests that people's political orientation shapes how they have understood and responded to the COVID-19 pandemic (Grossman et al. 2020). As a result, we analyzed whether respondents' partisanship affected their beliefs and behavioral responses after viewing a manipulated graph from the disinformation treatment. Shown in column 3 of Table A18, political orientation affects people's willingness to share the graph with their friends and followers on social media. The more conservative a

person's political orientation, the more likely they will share the graph (P = 0.022, t-test). In contrast, political orientation does not significantly affect any of the COVID-19 risk perceptions, behaviors, or trust in the media and public officials (see Tables A16, A17, and A18).

Table A16: Effect of Manipulated COVID-19 Graph on Risky Behaviors (by Political Orientation)

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. variable	Grocery	Restaurant	Concert	Visit a friend	Wear a mask	Travel
Disinformation	0.609***	0.449***	0.342***	0.391***	-0.044	0.493***
	(0.074)	(0.076)	(0.080)	(0.073)	(0.082)	(0.076)
Disinf. × Political	0.022	0.011	0.040	0.024	0.024	-0.012
	(0.039)	(0.040)	(0.040)	(0.039)	(0.041)	(0.040)
Age	-0.008***	-0.013***	-0.010***	-0.012***	0.006***	-0.012***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Female	-0.002	-0.072	-0.073	-0.135**	0.126*	-0.083
	(0.068)	(0.066)	(0.068)	(0.067)	(0.066)	(0.068)
Political	0.103***	0.175***	0.169***	0.150***	-0.233***	0.113***
	(0.029)	(0.028)	(0.030)	(0.028)	(0.030)	(0.028)
White	0.216***	0.220***	0.279***	0.307***	-0.284***	0.248***
	(0.077)	(0.072)	(0.069)	(0.075)	(0.068)	(0.074)
Constant	-0.018	0.353***	0.264**	0.337***	-0.269**	0.264**
	(0.124)	(0.115)	(0.115)	(0.121)	(0.126)	(0.122)
Observations	775	775	775	775	775	775
P-value corrected for MHT (Disinformation treatment)	1.000	1.000	1.000	1.000	1.000	1.000

Notes: OLS regressions with robust standard errors in parenthesis. OLS regressions with robust standard errors in parenthesis. All dependent variables were measured on a 7-point scale from 1 ("Very uncomfortable") to 7 ("Very comfortable"), and then standardized with mean 0 and standard deviation of 1. Disinformation is a dummy variable that has the value of 1 when subjects were shown the manipulated graph. "Age" is a continuous variable that measures a subject's age. "Female" is a dummy variable that has a value of 1 if a subject is female. "Political" is a variable that ranges from -3 to 3, and asks if a subject is politically liberal (-3) or conservative (3). "White" is a dummy that indicates if a subject's race is white. "Political x Disinformation" is the interaction term between political orientation and the disinformation job, indicating whether more conservative people respond differently to the manipulated graph. When we

control for background characteristics, the sample size decreases by nineteen because they did not report their age or ethnicity (or both). Significance levels: * p < 0.10, ** p < 0.05, *** p < 0.01. For the disinformation treatment, we additionally report p-values corrected for multiple hypothesis testing (MHT) using the Bonferroni correction method (six hypotheses).

Table A17: Effect of Manipulated COVID-19 Graph on Risk Perceptions (by Political Orientation)

	(7)	(8)	(9)
	Health	New	Vaccine
Dep. variable	concerns	variant	intent
Disinformation	-0.243***	-0.298***	-0.422***
	(0.077)	(0.076)	(0.077)
Disinf. × Political	-0.004	0.021	0.039
	(0.041)	(0.039)	(0.042)
Age	0.004*	0.008***	0.005**
	(0.002)	(0.002)	(0.002)
Female	0.005	0.093	-0.118*
	(0.071)	(0.068)	(0.069)
Political	-0.119***	-0.193***	-0.108***
	(0.032)	(0.029)	(0.029)
White	-0.311***	-0.273***	-0.234***
	(0.083)	(0.079)	(0.084)
Constant	0.081	-0.216*	0.164
	(0.132)	(0.124)	(0.132)
Observations	775	775	772
P-value corrected for MHT			
(Disinformation treatment)	1.000	1.000	0.357

Notes: OLS regressions with robust standard errors in parenthesis. The dependent variables "Health concerns" and "New variant" were measured on a 7-point scale from 1 ("Not at all worried") to 7 ("Extremely worried"). The dependent variable was measured on a 7-point scale from 1 ("Much less likely to be vaccinated") to 7 ("Much more likely to be vaccinated"). All dependent variables were standardized with mean 0 and standard deviation of 1. Disinformation is a dummy variable that has the value of 1 when subjects were shown the manipulated graph. "Age" is a continuous variable that measures a subject's age. "Female" is a dummy variable that has a value of 1 if a subject is female. "Political" is a

variable that ranges from -3 to 3, and asks if a subject is politically liberal (-3) or conservative (3). "White" is a dummy that indicates if a subject's race is white. "Political x Disinformation" is the interaction term between political orientation and the disinformation job, indicating whether more conservative people respond differently to the manipulated graph. When we control for background characteristics, the sample size in columns 1 and 2 decreases by nineteen because they did not report their age or ethnicity (or both). The sample size decreases by an additional three subjects in column 3 due to nonresponse to the vaccination question. Significance levels: * p < 0.10, ** p < 0.05, *** p < 0.01. For the disinformation treatment, we additionally report p-values corrected for multiple hypothesis testing (MHT) using the Bonferroni correction method (three hypotheses).

Table A18: Effect of Manipulated COVID-19 Graph on on Trust in Institutions and Media, and Sharing Intent (by Political Orientation)

Dep. variable	(10) Trust in media	(11) Trust in health officials	(13) Sharing Intent
Disinformation	-0.194***	0.012	-0.076
	(0.072)	(0.077)	(0.074)
Disinf. × Political	-0.030	-0.061	0.094**
	(0.038)	(0.042)	(0.041)
Age	0.008***	0.004*	-0.000
	(0.002)	(0.002)	(0.002)
Female	-0.153**	-0.022	-0.120*
	(0.064)	(0.066)	(0.071)
Political	-0.253***	-0.195***	-0.057*
	(0.028)	(0.030)	(0.030)
White	-0.239***	-0.178**	-0.414***
	(0.074)	(0.077)	(0.088)
Constant	-0.193	-0.215*	0.415***
	(0.119)	(0.125)	(0.134)
Observations	775	775	775
P-value corrected for MHT (Disinformation treatment)	1.000	1.000	0.066

Notes: OLS regressions with robust standard errors in parenthesis. The dependent variable "Trust in media" was measured on a 7-point scale from 1 ("No trust at all" to 7 ("A lot of trust"). The dependent variable "Trust in health officials" was measured on a 7-point scale from 1 ("Poor" to "Excellent"). The dependent variable "Sharing intent" was measured on a 7-point scale from 1 ("Extremely unlikely" to 7 ("Extremely likely"). All dependent variables were standardized with mean 0 and standard deviation of 1. Disinformation is a dummy variable that has the value of 1 when subjects were shown the manipulated graph. "Age" is a continuous variable that measures a subject's age. "Female" is a dummy variable that has a value of 1 if a subject is female. "Political" is a variable that ranges from -3 to 3, and asks if a subject is politically liberal (-3) or conservative (3). "White" is a dummy that indicates if a subject's race is white. "Political x Disinformation" is the interaction term between political orientation and the disinformation job, indicating whether more conservative people respond differently to the manipulated graph. When we control for background characteristics, the sample size decreases by nineteen because they did not report their age or ethnicity (or both). Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For the disinformation treatment, we additionally report p-values corrected for multiple hypothesis testing (MHT) using the Bonferroni correction method (two hypotheses and one hypothesis).

3.4 Platform Intervention

As pre-registered, Table A19 reports the regression results for the platform intervention survey. Column 1 reports forecasters' predictions about how each intervention will affect how many workers (out of 100) would be willing to complete the disinformation job. Unlike the non-parametric analysis in the main text, here we control for forecasters' background characteristics. However, the results remain nearly identical to those reported in the main text. Forecasters predicted the "Accountability" treatment would be the most effective deterrent for inhibiting workers from completing the disinformation job ($P < 0.001$, t-test adjusted for MHT). The remaining interventions ("Reminder," "Training," "Whistleblowing," and "Peer Information") are predicted to have similar effects on workers' estimates on willingness to complete the disinformation job ($P < 0.001$, t-test, post-hoc analysis), though these interventions are predicted to be half as effective as the Accountability treatment (all *P*-values are smaller than 0.001, t-tests adjusted for MHT).

We also asked the forecasters to report their level of confidence about their predictions and split the sample into those with high confidence (above-median confidence score) and those with low confidence (below-median confidence score). This allows us to check whether forecasters' predictions about the relative effectiveness of the interventions are impacted by their confidence, particularly whether the results change when accounting for the uncertainty in the predictions of those who had lower confidence. Column 2 presents the results for the high-confidence forecasters and column 3 for the low-confidence forecasters, respectively. Those with lower-confidence predict the suspension treatment to have a bigger effect compared to those with higher-confidence ($P < 0.001$, t-test, post-hoc analysis), though both groups believe the intervention to be the most effective at discouraging workers from completing the disinformation job.

Lower-confidence forecasters tend to predict that the interventions will be more effective than those with higher-confidence (*P*-values range from 0.005 to 0.688, t-tests, post-hoc analysis). However, the results are qualitatively similar for both groups.

We randomized the order of the interventions for each participant. Table A20 examines the order effects. As a benchmark, column (1) reproduces the regression results without accounting for order effects. Column (2) shows that the longer people consider potential interventions, the more likely they believe that an intervention is effective at preventing people from working on the disinformation job ($P < 0.001$, t-test, post-hoc analysis). However, controlling for order effects does not change the rank order of the interventions' effectiveness. Column (3) shows the results for when we limit the analysis to the first intervention that forecasters were presented with, after having seen the baseline. As shown, the results remain qualitatively the same. One exception is the whistleblowing treatment; its coefficient is smaller and not significant anymore ($P = 0.203$, t-test, post-hoc analysis).

Table A19: Predicted Effectiveness of Platform Interventions

	(1)	(2)	(3)
	Predicted worker acceptance for disinformation		
Dep. variable	job (0-100)		
Accountability	-21.335***	-14.232***	-26.271***
	(1.537)	(2.325)	(1.994)
Reminder	-10.275***	-8.305***	-11.644***
	(1.341)	(2.008)	(1.800)
Training	-10.713***	-9.396***	-11.627***
	(1.290)	(1.808)	(1.797)
Whistleblowing	-12.470***	-13.207***	-11.958***
	(1.510)	(2.494)	(1.896)
Peer Information	-12.865***	-9.037***	-15.525***
	(1.202)	(1.592)	(1.697)
Age	-0.067	0.157	-0.155
	(0.107)	(0.166)	(0.126)
Female	1.050	-2.354	3.095
	(2.123)	(3.089)	(2.651)
Full-time	10.071***	10.817	7.370**
	(3.474)	(7.250)	(3.733)
Income	5.062***	4.524***	3.033**
	(1.058)	(1.608)	(1.399)
Political	1.222*	1.146	0.480
	(0.684)	(0.866)	(0.994)
Education:			
Some college	-6.783	13.490	-13.635**
	(6.212)	(14.749)	(6.259)
College degree	-0.288	17.998	-6.705
	(5.586)	(12.611)	(5.632)
Master's/PhD	0.019	17.821	-6.069
	(5.941)	(12.890)	(6.071)

Constant	43.005***	24.363*	58.710***
	(7.853)	(12.925)	(9.830)
Observations	2,400	984	1,416
Sample	Full		
		Above-median confidence	Below-median confidence
Largest p-value (corrected for MHT, five hypotheses)	<0.001	<0.001	<0.001

Notes: OLS regressions with standard errors clustered at the individual level in parentheses. The dependent variable is the number of workers predicted to work on the disinformation job. "Accountability," "Reminder," "Training," "Whistleblowing" and "Peer Information" are dummy variables for the treatments. "Age" is a continuous variable that measures a subject's age. "Female" is a dummy variable that has a value of 1 if a subject is female. "Full-time" is a dummy variable that measures if a subject is employed full-time in their job. "Income" is a variable that ranges from 1 to 7, where a subject is asked to report its household income, compared to the average income in the US. "Political" is a variable that ranges from -3 to 3, and asks if a subject is politically liberal (-3) or conservative (3). "High school," "Some college," "College," and "Master's/PhD" are dummy variables for a subject's highest level of education. Due to very few observations with a Master's degree or higher, we combine "Master's" and "PhDs" into one category. Column 1 reports the results for the full sample; columns 2 and 3 present the results by subjects' level of confidence (above- vs. below-median). Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. At the bottom, we additionally report p-values corrected for MHT (five hypotheses) against the null of no change relative to the status quo.

Table A20: Predicted Effectiveness of Platform Interventions (Controlling for Order Effects)

	(1)	(2)	(3)
Dep. variable	Predicted worker acceptance for disinformation job (0-100)		
Accountability	-21.335***	-17.040***	-16.912***
	(1.537)	(1.660)	(3.315)
Reminder	-10.275***	-6.375***	-7.113**
	(1.341)	(1.439)	(2.929)
Training	-10.713***	-6.710***	-10.447***
	(1.290)	(1.457)	(2.715)
Whistleblowing	-12.470***	-8.458***	-3.554
	(1.510)	(1.720)	(2.788)
Peer Information	-12.865***	-8.655***	-10.375***
	(1.202)	(1.409)	(2.787)
Age	-0.067	-0.067	-0.065
	(0.107)	(0.107)	(0.111)
Female	1.050	1.050	-0.356
	(2.123)	(2.124)	(2.127)
Full-time	10.071***	10.071***	5.005
	(3.474)	(3.475)	(3.453)
Income	5.062***	5.062***	4.006***
	(1.058)	(1.058)	(1.099)
Political	1.222*	1.222*	0.577
	(0.684)	(0.685)	(0.703)
Education:			
Some college	-6.783	-6.783	1.350
	(6.212)	(6.213)	(6.196)
College degree	-0.288	-0.288	5.909
	(5.586)	(5.587)	(5.692)
Master's/PhD	0.019	0.019	2.857
	(5.941)	(5.942)	(6.077)

Order of treatment		-1.361***	
		(0.329)	
Constant	43.005***	44.367***	47.232***
	(7.853)	(7.849)	(8.051)
Observations	2,400	2,400	800
Rounds	All	All	First treatment only
Largest p-value (corrected for MHT, five hypotheses)	<0.001	<0.001	1.000

Notes: OLS regressions with standard errors clustered at the individual level in parentheses. The dependent variable is the number of workers predicted to work on the disinformation job. "Accountability," "Reminder," "Training," "Whistleblowing" and "Peer Information" are dummy variables for the treatments. "Age" is a continuous variable that measures a subject's age. "Female" is a dummy variable that has a value of 1 if a subject is female. "Full-time" is a dummy variable that measures if a subject is employed full-time in their job. "Income" is a variable that ranges from 1 to 7, where a subject is asked to report its household income, compared to the average income in the US. "Political" is a variable that ranges from -3 to 3, and asks if a subject is politically liberal (-3) or conservative (3). "High school," "Some college," "College," and "Master's/PhD" are dummy variables for a subject's highest level of education. Due to very few observations with a Master's degree or higher, we combine "Master's" and "PhDs" into one category. "Order of treatment" is a count variable (1-6) for the order in which the treatment was presented to each forecaster. Column 1 reproduces the main result from column 1, Table A19. Column 2 controls for order effects, and column 3 restricts the analysis to the predicted effectiveness of the first randomly shown intervention. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. At the bottom, we additionally report p-values corrected for MHT (five hypotheses) against the null of no change relative to the status quo.

Table A21: P-values from Nonparametric Tests Corrected for MHT (Platform Intervention Survey)

	P-value (uncorrected)	P-value corrected for MHT (Bonferroni)
Accountability	<0.001	<0.001
Reminder	<0.001	<0.001
Training	<0.001	<0.001
Whistleblowing	<0.001	<0.001
Peer Information	<0.001	<0.001

Notes: The first column shows uncorrected p-values and the second column shows p-values corrected for multiple hypothesis testing using the Bonferroni method (five hypotheses).

4.1 Instructions for the field experiment

First job instructions

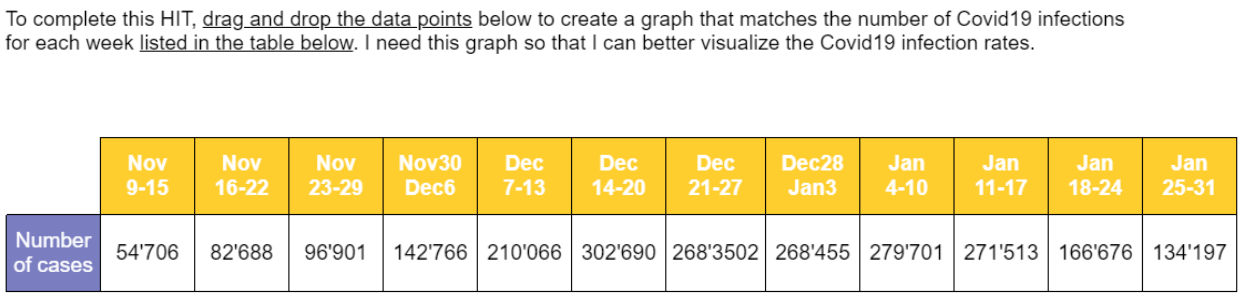
Here we show the instructions of the first job we posted on MTurk.

[Title] Simple Data Visualization

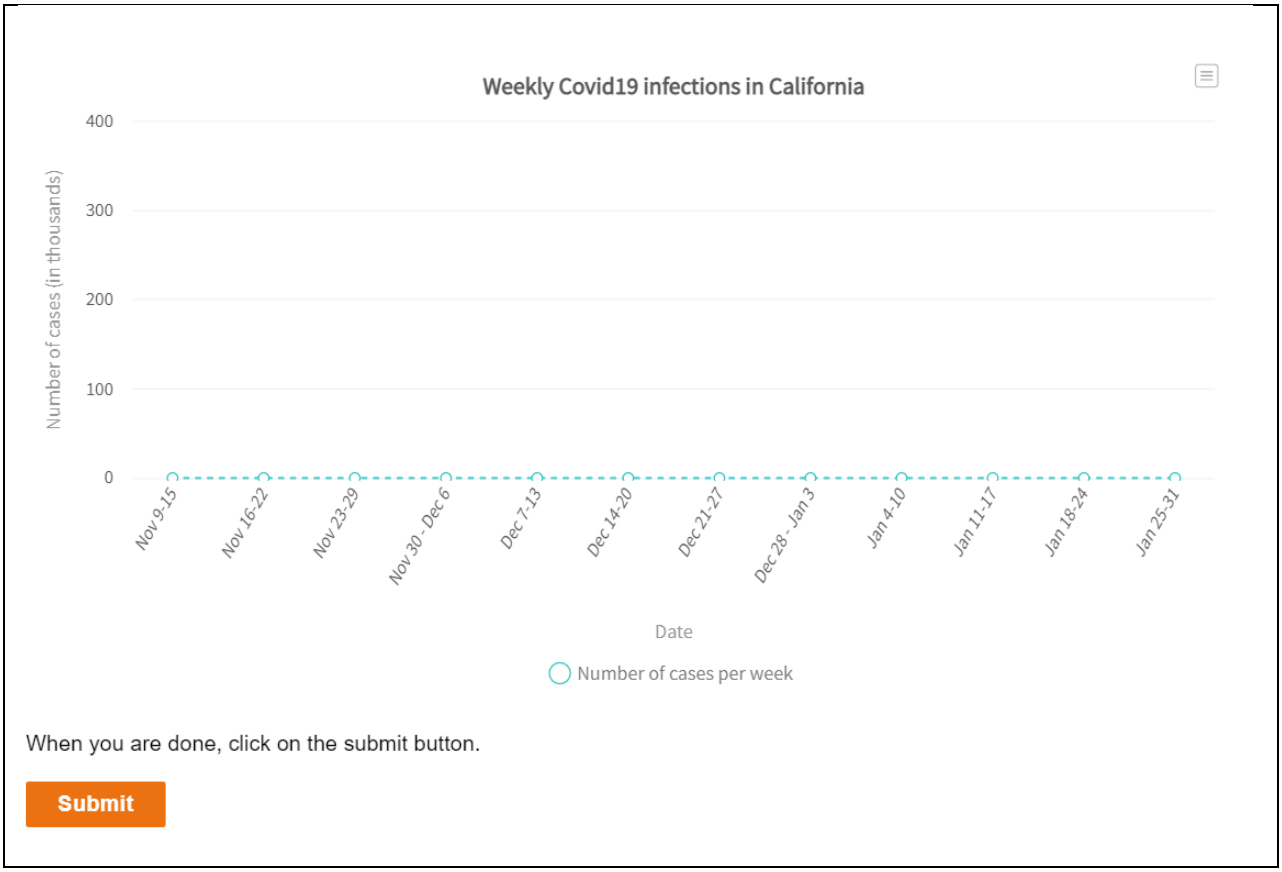
[Description] Create a graph by dragging and dropping data points to match the number of Covid19 infection rates for each week listed in a provided table.

[HIT instructions] To complete this HIT, drag and drop the data points below to create a graph that matches the number of Covid19 infections for each week listed in the table below. I need this graph so that I can better visualize the Covid19 infection rates.

Screenshot first job


To complete this HIT, drag and drop the data points below to create a graph that matches the number of Covid19 infections for each week listed in the table below. I need this graph so that I can better visualize the Covid19 infection rates.

	Nov 9-15	Nov 16-22	Nov 23-29	Nov30 Dec6	Dec 7-13	Dec 14-20	Dec 21-27	Dec28 Jan3	Jan 4-10	Jan 11-17	Jan 18-24	Jan 25-31
Number of cases	54'706	82'688	96'901	142'766	210'066	302'690	268'3502	268'455	279'701	271'513	166'676	134'197



Second job invitations

After completing the first stage, we randomly sent one of the following messages to our workers, to generate interest for the second job:

CONTROL TREATMENT INVITATION

[Header] Simple Data Visualization (Follow-Up) for $1.20

[Text] You recently completed a HIT visualizing Covid19 infections. I'm inviting you for a follow-up HIT to create a similar graph for Covid19 deaths. The HIT will pay $1.20 and it will again take about 6 minutes to complete.

I will publicly post this new graph on Facebook and Twitter.

Please search for “Michael Miller” or “Simple Data Visualization (Follow-Up)” to accept the HIT.

LOWER-PAY TREATMENT INVITATION

[Header] Simple Data Visualization (Follow-Up) for $0.60

[Text] You recently completed a HIT visualizing Covid19 infections. I’m inviting you for a follow-up HIT to create a similar graph for Covid19 deaths. The HIT will pay $0.60 and it will again take about 6 minutes to complete.

I will publicly post this new graph on Facebook and Twitter.

Please search for “Michael Miller” or “Simple Data Visualization (Follow-Up)” to accept the HIT.

DISINFORMATION TREATMENT INVITATION

[Header] Simple Data Visualization (Follow-Up) for $1.20

[Text] You recently completed a HIT visualizing Covid19 infections. I’m inviting you for a follow-up HIT to create a similar graph for Covid19 deaths, but this time I need you to reduce the numbers compared to what the “original” numbers would look like. The HIT will pay $1.20 and it will again take about 6 minutes to complete.

When you are done, the Covid19 death curve must look much flatter than the original entries. I don’t care how you change the numbers as long as someone can tell that the Covid19 death rate looks less worrying after viewing the manipulated graph.

I will publicly post this new graph on Facebook and Twitter.

Please search for “Michael Miller” or “Simple Data Visualization (Follow-Up)” to accept the HIT.

Second job instructions

Workers who were interested in our second job could then look into more detail on what the job entailed, before accepting and working on this job. They would see the following information:

CONTROL TREATMENT second job instructions (This is the same as the Lower-pay treatment.)

[Title] Simple Data Visualization (Follow up)

[Description] Create a graph by dragging and dropping data points to match the number of Covid19 death rates for each week listed in a provided table.

[HIT instructions] To complete this HIT, drag and drop the data points to create a graph that matches the number of Covid19 deaths for each week listed in the table below.

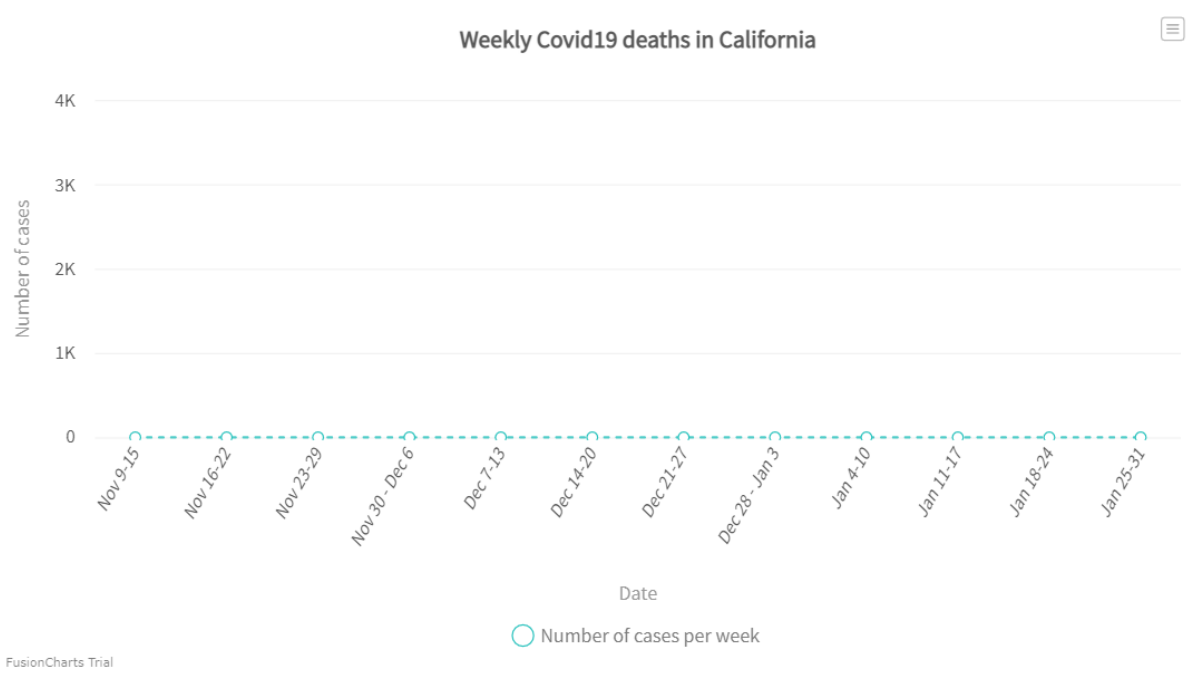
I will publicly post this new graph on Facebook and Twitter.

CONTROL TREATMENT Screenshot second job (This is the same as the Lower-pay treatment.)

To complete this HIT, drag and drop the data points to create a graph that matches the number of Covid19 deaths for each week listed in the table below.

I plan to publicly post this new graph on Facebook and Twitter.

	Nov 9-15	Nov 16-22	Nov 23-29	Nov30 Dec6	Dec 7-13	Dec 14-20	Dec 21-27	Dec28 Jan3	Jan 4-10	Jan 11-17	Jan 18-24	Jan 25-31
Number of cases	290	423	445	755	1093	1624	1627	2318	3163	3691	3398	3907



When you are done, click on the submit button.

Submit

DISINFORMATION TREATMENT Second job instructions

[Title] Simple Data Visualization (Follow up)

[Description] Create a graph by dragging and dropping data points such that the Covid19 death rate looks less worrying for each week listed in a provided table.

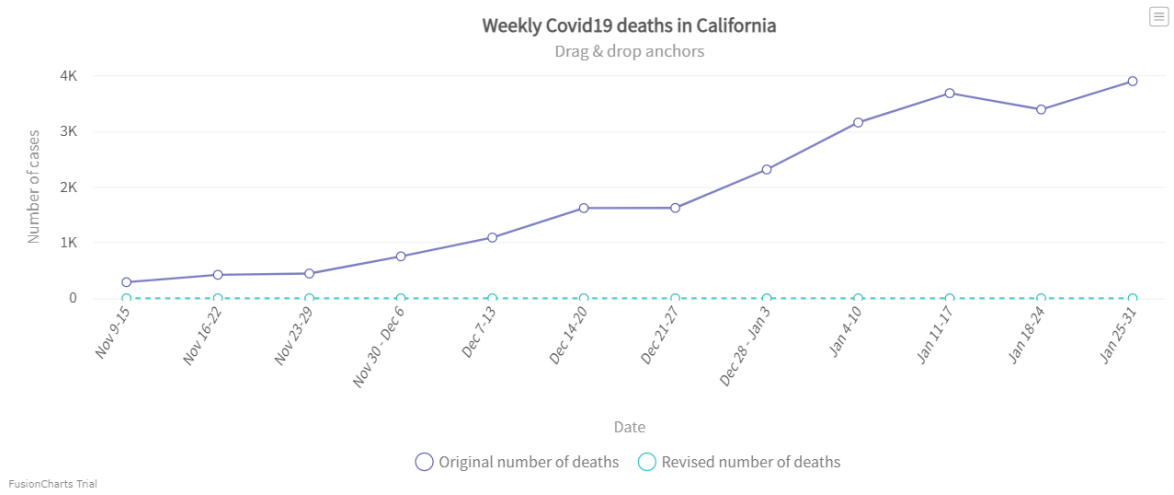
[HIT instructions]
To complete this HIT, drag and drop the data points to create a graph again, but this time reduce the numbers of Covid19 deaths (especially after Dec 7-13) so that the manipulated turquoise line looks much flatter than the original purple line.

I don't really care how you change the turquoise line as long as someone looking at the manipulated line can tell that the Covid19 death rate looks less worrying than the original curve.

I will publicly post this new graph on Facebook and Twitter.

DISINFORMATION TREATMENT Screenshot second job

	Nov 9-15	Nov 16-22	Nov 23-29	Nov30 Dec6	Dec 7-13	Dec 14-20	Dec 21-27	Dec28 Jan3	Jan 4-10	Jan 11-17	Jan 18-24	Jan 25-31
Revised number of deaths	0	0	0	0	0	0	0	0	0	0	0	0
Original number of deaths	290	423	445	755	1093	1624	1627	2318	3163	3691	3398	3907



When you are done, click on the submit button.

Submit

4.2 Survey instructions for the Manipulation check study

Description

This study is being conducted by researchers at Erasmus University Rotterdam, the University of Michigan, and Northwestern University. We are interested in your views as a worker on MTurk.

Duration

It should take 4 minutes or less to complete the survey.

Compensation

For your participation, you will be paid $1. This survey contains two questions that will check whether you pay attention to the instructions. If you do not answer them correctly, the survey ends and you will not receive the $1 participation fee.

Risks and Benefits

This study does not involve any known physical or emotional risk. Beyond the payment you may receive, your participation contributes to the advance of scientific knowledge.

Confidentiality

The information collected may be published in scientific journals or academic presentations, but your personal identity or involvement as a participant will not be revealed. We process personal data in accordance with the EU General Data Protection Regulation (GDPR). Information collected during this study will be retained by the researchers, but any data that could identify you will be deleted after completion of the study.

Subject's Rights

Your participation is voluntary and you may quit at any time without any negative consequences, except that you will not receive the $1 participation fee.

If you have questions about this project, you may contact Jan Stoop at: stoop@ese.eur.nl.

Thank you for your participation!

Please check the box below to confirm that you are at least 18 years old, have read and understood this consent form, and agree to participate in this study.

I'm at least 18, understand the consent form, and agree to participate.

I'm not 18 and/or I don't agree.

This study should take about 4 minutes to complete. It is important to take your time to read all instructions and questions carefully before you answer them. Previous research has found that some people do not take the time to read everything that is displayed in a survey. The questions below test whether you are able to read and recall portions of the survey. Therefore, please answer 'five' on the first question, subtract 'three' from this number and use the result as the answer on the second question.

In general, I would rather read a book than watch a movie:

0: Strongly disagree

1

2

3

4

5

6: Strongly agree

I like reading e-books (for example, on a Kindle or an iPad):

0: Strongly disagree

1

2

3

4

5

6: Strongly agree

We are interested in your views as an MTurker. Imagine you recently completed a HIT that asked you to graph US data on Covid19 infection rates. After completing the HIT, the same requester invites you back to do a follow-up HIT. This follow-up HIT pays $1.20 and takes about 6 minutes to complete.

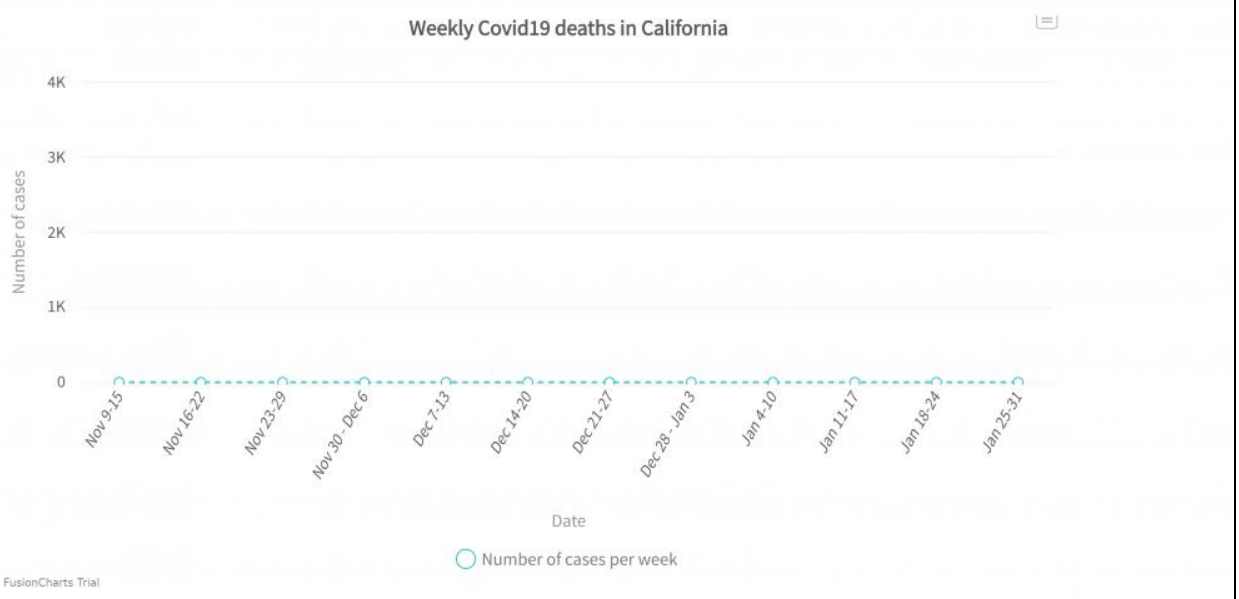
On the next page you will see this follow-up HIT. Please read it carefully before you continue the survey.

JOB DESCRIPTION FOR THE CONTROL TREATMENT:

To complete this HIT, drag and drop the data points to create a graph that matches the number of Covid19 deaths for each week listed in the table below.

I will publicly post this new graph on Facebook and Twitter.

	Nov 9-15	Nov 16-22	Nov 23-29	Nov30 Dec6	Dec 7-13	Dec 14-20	Dec 21-27	Dec28 Jan3	Jan 4-10	Jan 11-17	Jan 18-24	Jan 25-31
Number of cases	54'706	82'688	96'901	142'766	210'066	302'690	268'3502	268'455	279'701	271'513	166'676	134'197



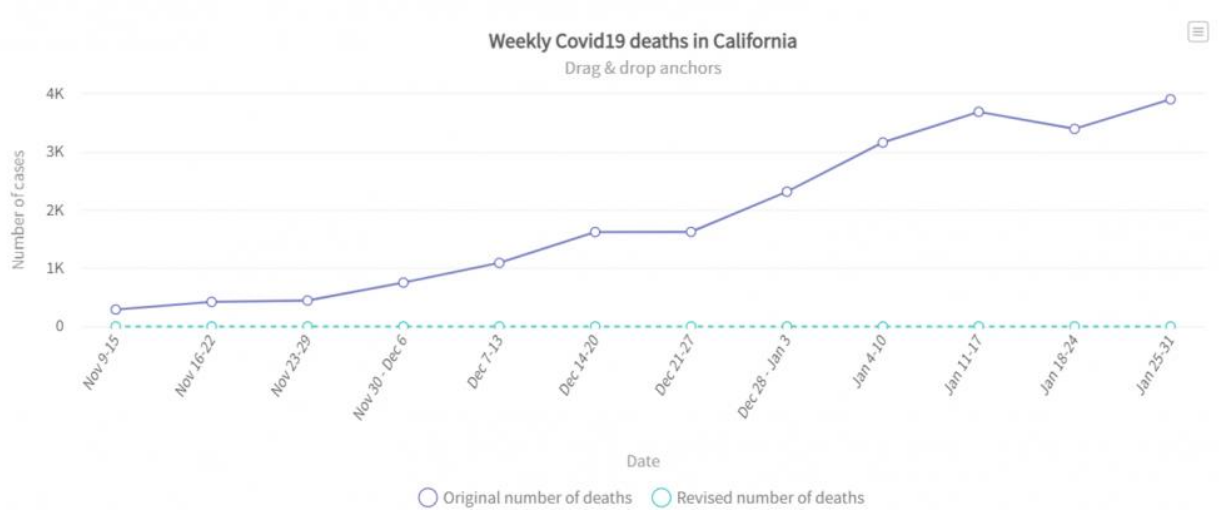
JOB DESCRIPTION FOR THE DISINFORMATION TREATMENT:

To complete this HIT, drag and drop the data points to create a graph again, but this time reduce the numbers of Covid19 deaths (especially after Dec 7-13) so that the manipulated turquoise line looks much flatter than the original purple line.

I don't really care how you change the turquoise line as long as someone looking at the manipulated line can tell that the Covid19 death rate looks less worrying than the original curve.

I will publicly post this new graph on Facebook and Twitter.

	Nov 9-15	Nov 16-22	Nov 23-29	Nov30 Dec6	Dec 7-13	Dec 14-20	Dec 21-27	Dec28 Jan3	Jan 4-10	Jan 11-17	Jan 18-24	Jan 25-31
Revised number of deaths	0	0	0	0	0	0	0	0	0	0	0	0
Original number of deaths	290	423	445	755	1093	1624	1627	2318	3163	3691	3398	3907



Do you personally believe that working on this follow-up HIT is morally acceptable, morally unacceptable, or is it not a moral issue?

Morally acceptable

Morally unacceptable

Not a moral issue

Would you accept and complete this follow-up HIT?

Yes

No

I'm not sure

What is the lowest pay you would accept to work on this follow-up HIT (recall, it takes about 6 minutes to complete the HIT)? The slider is set in US dollars.

0 1 1 2 2 3 3 4 4 5

Have you ever encountered a HIT asking you to fabricate or manipulate data in a misleading way?

Yes

No

Please describe the HIT that asked you to fabricate or manipulate data in a misleading way.

_____

Have you ever decided not to complete a HIT because you found it morally unacceptable?

Yes

No

Please describe the HIT that you found morally unacceptable.

_____

Suppose the follow-up HIT is being offered to MTurkers from the U.S. (with an approval rate of at least 95%; minimum of 500 HITs completed). Out of 100 MTurkers, how many do you think would accept and work on the follow-up HIT?

0 10 20 … … 90 100

Suppose **you** completed the follow-up HIT, and created the graph as requested, and it was posted on social media (for example, on Facebook). To what extent would **you** feel personally responsible for how it affects other people?

	Not at all responsible						Fully responsible
	1	2	3	4	5	6	7
							

Please write for each statement how much you agree with it:

	Not true (1)	(2)	(3)	Somewhat true (4)	(5)	(6)	Very true (7)
I sometimes tell lies if I have to.							
I never cover up my mistakes.							
There have been occasions where I have taken advantage of someone.							
I sometimes try to get even rather than to forgive and forget.							
I have said something bad about a friend behind his or her back.							
When I hear people talk privately, I avoid listening.							
I never take things that don't belong to me.							

I don't gossip about other people's business.

Now please tell us about yourself:

What is your age?

_____

What is your gender?

Male

Female

Other

Which category best describes your highest level of education?

High school / GED or less

Some college

College degree

Master's or professional degree (for example JD, MD, MBA)

Doctoral degree

What is your current employment status?

Full-time employee

Part-time employee

Self-employed or small business owner

Unemployed and looking for work

Student

Not in labor force (for example: retired, or full-time parent)

What is your household income compared to the average household income in the U.S.?

Much lower than average income						Much higher than average income
1	2	3	4	5	6	7

In general, to what extent are you politically liberal or conservative?

Very liberal						Very conservative
1	2	3	4	5	6	7

Who did you vote for in the 2020 elections?

- Trump
- Biden
- Other
- I did not vote

In which state do you currently reside?

▼ Alabama (1) ... I do not reside in the United States (53)

Do you have any other comments of suggestions that you would like to share with us? Is there anything that is unclear or confusing? Please let us know what you think.

_____

We thank you for your time spent taking this survey.

Your response has been recorded.

4.3 Survey instructions for the Downstream Consequences study

Principal Investigators: XXX **IRB Study Number**: HUM00179761

Description

This study is being conducted by researchers at the University of Michigan, Northwestern University, and Erasmus University Rotterdam. We are interested in your views about the Covid19 pandemic.

Duration

It should take 5 minutes or less to complete the survey.

Compensation

For your participation, you will be paid 0.63 pounds (about US $0.87).

Risks and Benefits

This study does not involve any known physical or emotional risk. Beyond the payment you may receive, your participation contributes to the advance of scientific knowledge.

Confidentiality

The information collected may be published in scientific journals or academic presentations, but your personal identity or involvement as a participant will not be revealed. We process personal data in accordance with the EU General Data Protection Regulation (GDPR). Information collected during this study will be retained by the researchers, but any data that could identify you will be deleted after completion of the study.

Subject's Rights

Your participation is voluntary and you may quit at any time without any negative consequences, except that you will not receive the participation fee. If you have any questions or concerns regarding this study, its purpose or procedures, or if you have a research-related problem, please feel free to contact Jan Stoop at: stoop@ese.eur.nl. If you have any questions concerning your rights as a research participant, you may contact the University of Michigan Institutional Review Board office by calling 734-936-0933 or

emailing irbhsbs@umich.edu.

Thank you for your participation!

Please check the box below to confirm that you are at least 18 years old, have read and understood this consent form, and agree to participate in this study.

☐ I'm at least 18, understand the consent form, and agree to participate.

☐ I'm not 18 and/or I don't agree.

Please copy your Prolific ID in the text box below:

_____

This study should take about 5 minutes or less to complete. It is important to take your time to read all instructions and questions carefully before you answer them. Previous research has found that some people do not take the time to read everything that is displayed in a survey. The questions below test whether you are able to read and recall portions of the survey. Therefore, please answer 'five' on the first question, subtract 'three' from this number and use the result as the answer on the second question.

In general, I would rather read a book than watch a movie:

0: Strongly disagree

1

2

3

4

5

6: Strongly agree

I like reading e-books (for example, on a Kindle or an iPad):

0: Strongly disagree

1

2

3

4

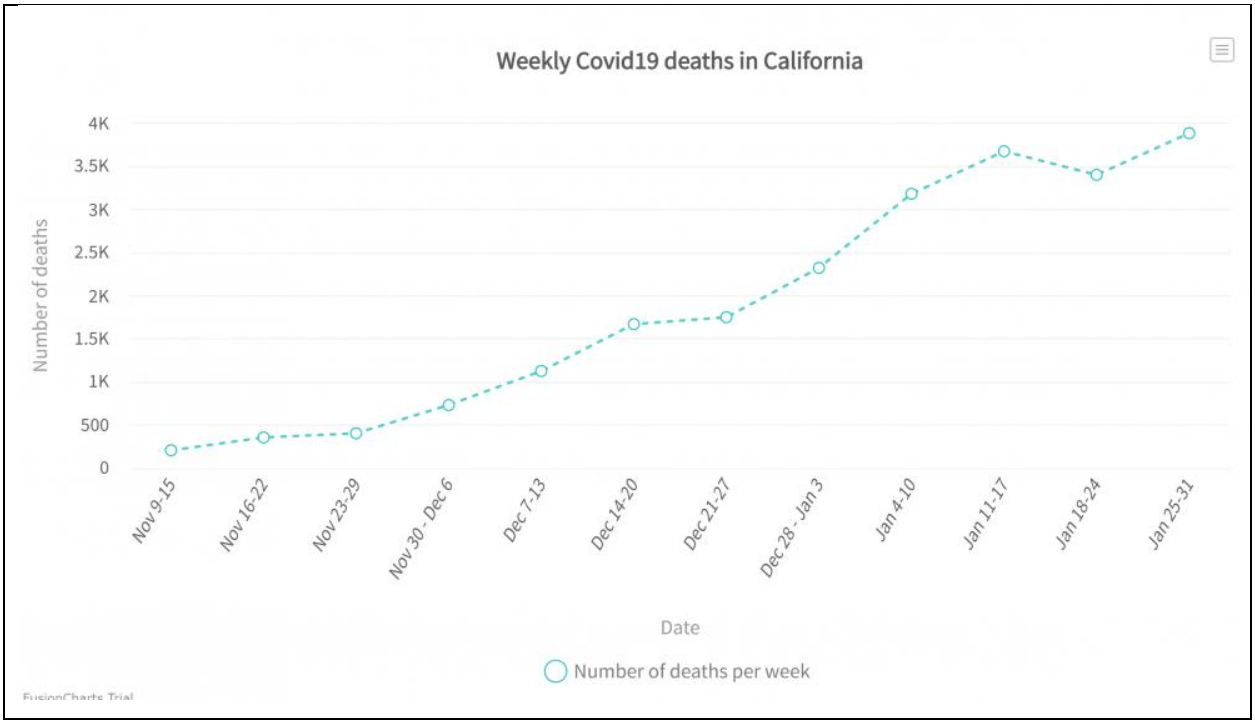
5

6: Strongly agree

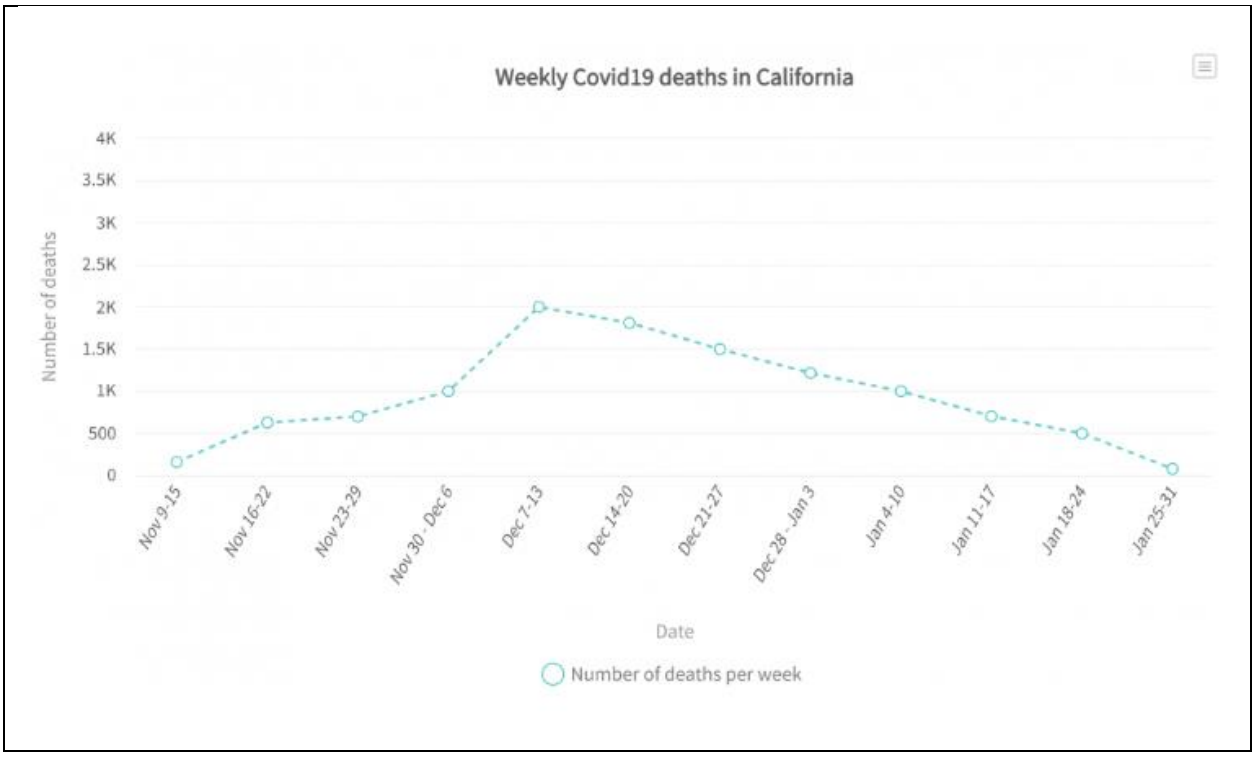
In this survey, we will present you with some Covid19 data, and we will then ask some questions on how you react to the information that you are given.

Suppose you see the following graph posted on social media (for example on Facebook or Twitter).

CONTROL TREATMENT



DISINFORMATION TREATMENT



After seeing this graph, how likely would you be to share this graph with your friends and/or followers on social media?

	Extremely unlikely						Extremely likely
	1	2	3	4	5	6	7

After seeing this graph, would you feel comfortable or uncomfortable doing each of the following in California?

	Very uncomfortable 1 2 3 4 5 6 7 Very comfortable
Going out to the grocery store	
Eating out in a restaurant	
Attending an indoor sporting event or concerts	
Visiting with a close friend or family member inside their home	
Support mandatory mask wearing on public places in California	
Travel to California for a pre-paid trip in the next month	

After seeing this graph, how worried are you about the health consequences of Covid19 for you?

	Not at all worried 1 2 3 4 5 6 7 Extremely worried
	

After seeing this graph, how worried are you that the Covid19 mutation will lead to a new wave of infections?

	Not at all worried					Extremely worried	
	1	2	3	4	5	6	7

Overall, the information provided in this graph makes me

	Much less likely to be vaccinated			No less or more likely to be vaccinated			Much more likely to be vaccinated
	1	2	3	4	5	6	7

Q114 Overall, how likely are you to fact-check the information in this graph via other sources?

Extremely likely

Extremely unlikely

1 2 3 4 5 6 7

Have you seen similar content online in the last month on social media?

- Yes
- No
- Do not know

After seeing this graph, how would you rate your trust in the mainstream media's reporting of Covid19?

No trust at all A lot of trust

1 2 3 4 5 6 7

After seeing this graph, how would you rate your trust in the job Public health officials, such as those at the CDC (Centers for Disease Control and Prevention), are doing responding to the Covid19 outbreak?

Poor Excellent

1 2 3 4 5 6 7

Finally, we have some questions about you.

Have you already received a Covid19 vaccine?

Yes

No

What sources of information do you trust regarding Covid19? (Please choose all that apply.)

Television news

Newspaper and other journalism

government briefings or websites

National health authorities

Scientific experts

Social media platforms (e.g. Facebook, Twitter, Youtube)

Celebrities

Online search engines or other websites (e.g. Google)

Family and friends

Work/School/College guidelines

None of the above

Other:

Please specify:

Which of the following mainstream media do you watch\follow mostly? (Maximum of 3 answers.)

ABC

CBS

CNN

NBC / MSNBC

Fox News

OAN (One America News)

Local News

NPR (Public Radio)

Huff Post

The New York Times

The Wall Street Journal

Washington Post

Breitbart

Other (multiple possible, separate with ",")

I don't follow any news

Please write down what other mainstream media you follow:

_____

In general, to what extent are you politically liberal or conservative?

	Very liberal						Very conservative
	1	2	3	4	5	6	7

Who did you vote for in the 2020 elections?

Trump

Biden

Other

I did not vote

In which state do you currently reside?

▼ Alabama (1) ... I do not reside in the United States (53)

Debriefing Information

Title of Research Study: The Supply of Misinformation about Covid19 Data
Principal Investigator: Jan Stoop

Thank you for your participation. The information below provides more details about the purpose of the study and data confidentiality.

Study Purpose:

We are interested in how people respond to data visualizations of health information. For this study, we asked workers to manipulate graphs about Covid19 deaths and now we want to see how people would react to it, if they would see the graphs on social media. Some participants saw the offical data, while others saw a graph that downplayed the severity of the Covid19 outbreak. The findings of this study will help us to better understand how we can mitigate the spread of misinformation, especially with regards to health information.

For your reference, here is a graph displaying the actual data about Covid19 deaths (source: ourworldindata.org).

Confidentiality:

As noted in the beginning of the study, the information collected in this survey may be published in scientific journals or academic presentations, but your personal identity or involvement as a participant will not be revealed. We process personal data in accordance with the EU General Data Protection Regulation (GDPR). Information collected during this study will be retained by the researchers, but any data that could identify you will be deleted after completion of the study.

Please do not disclose research procedures and/or hypotheses to anyone who might participate in this study in the future as this could affect the results of our research.

Contact Information:

Now that you are fully informed about the study purpose, you may decide that you do not want the data we collected used in this research. If you would like your anonymized data removed from the study and permanently deleted, please email stoop@ese.eur.nl within 30 days of receiving this message. If you would like your anonymized data removed from the study and permanently deleted, you can still keep the wages we paid for your participation.

If you have any questions or concerns regarding this study, its purpose or procedures, please feel free to contact Professor Jan Stoop at: stoop@ese.eur.nl.

If you want to see the results of this study, please let us know by email. We will send you a summary of the results.

We thank you for your time spent taking this survey.

Your response has been recorded.

4.4 Survey instructions for the Platform Interventions study

IRB Study Number: HUM00179761

Description

This study is being conducted by researchers at Erasmus University Rotterdam, the University of Michigan, and Northwestern University. We are interested in your views as a worker on MTurk.

Duration

It should take 5 minutes or less to complete the survey.

Compensation

For your participation, you will be paid $0.75. This survey contains two questions that will check whether you pay attention to the instructions. If you do not answer them correctly, the survey ends and you will not receive the $0.75 participation fee.

Risks and Benefits

This study does not involve any known physical or emotional risk. Beyond the payment you may receive, your participation contributes to the advance of scientific knowledge.

Confidentiality

The information collected may be published in scientific journals or academic presentations, but your personal identity or involvement as a participant will not be revealed. We process personal data in accordance with the EU General Data Protection Regulation (GDPR). Information collected during this study will be retained by the researchers, but any data that could identify you will be deleted after completion of the study.

Subject's Rights

Your participation is voluntary and you may quit at any time without any negative consequences, except that you will not receive the $0.75 participation fee.

If you have questions about this project, you may contact Jan Stoop at: stoop@ese.eur.nl. If you have any questions concerning your rights as a research participant, you may contact the University of Michigan Institutional Review Board office by calling 734-936-0933 or emailing irbhsbs@umich.edu.

Thank you for your participation!

Please check the box below to confirm that you are at least 18 years old, have read and understood this consent form, and agree to participate in this study.

I'm at least 18, understand the consent form, and agree to participate. (4)

I'm not 18 and/or I don't agree. (5)

ID Please write down your MTurk ID below, for payment related issues:

_____

This study should take about 5 minutes to complete. It is important to take your time to read all instructions and questions carefully before you answer them. Previous research has found that some people do not take the time to read everything that is displayed in a survey. The questions below test whether you are able to read and recall portions of the survey. Therefore, please answer 'five' on the first question, subtract 'three' from this number and use the result as the answer on the second question.

In general, I would rather read a book than watch a movie:

0: Strongly disagree

1

2

3

4

5

6: Strongly agree

I like reading e-books (for example, on a Kindle or an iPad):

0: Strongly disagree

1

2

3

4

5

6: Strongly agree

We are interested in your views as an MTurker. Imagine you recently completed a HIT that asked you to graph US data on Covid19 infection rates. After completing the HIT, the same requester invites you back to do a follow-up HIT. This follow-up HIT pays $1.20 and takes about 6 minutes to complete.

On the next page you will see this follow-up HIT. Please read it carefully before you continue the survey.

This is the job we want you to think about:

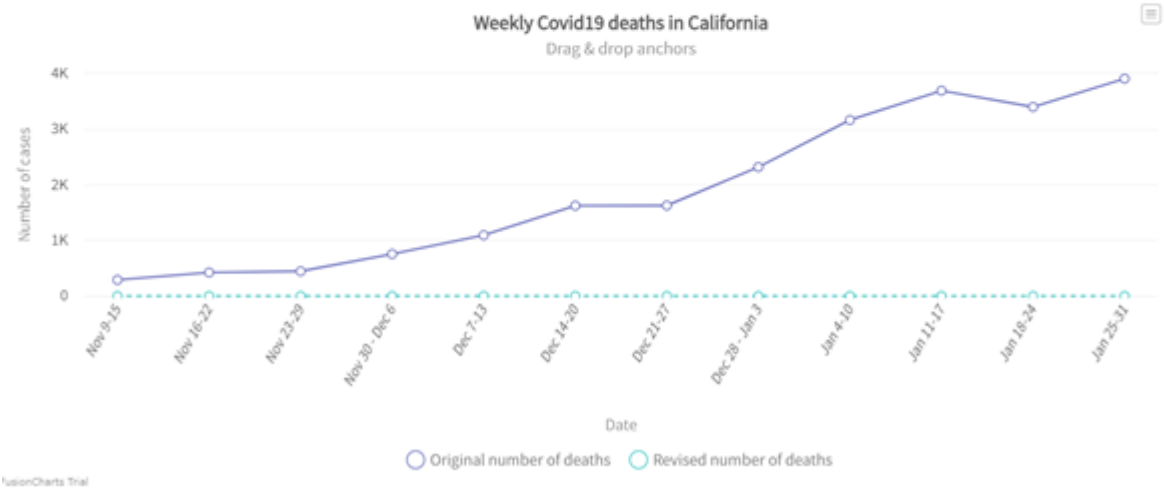
JOB DESCRIPTION:

To complete this HIT, drag and drop the data points to create a graph again, but this time reduce the number of Covid19 deaths (especially after Dec 7-13) so that the manipulated turquoise line looks much flatter than the original purple line.

I don't really care how you change the turquoise line as long as someone looking at the manipulated line can tell that the Covid19 death rate looks less worrying than the original curve.

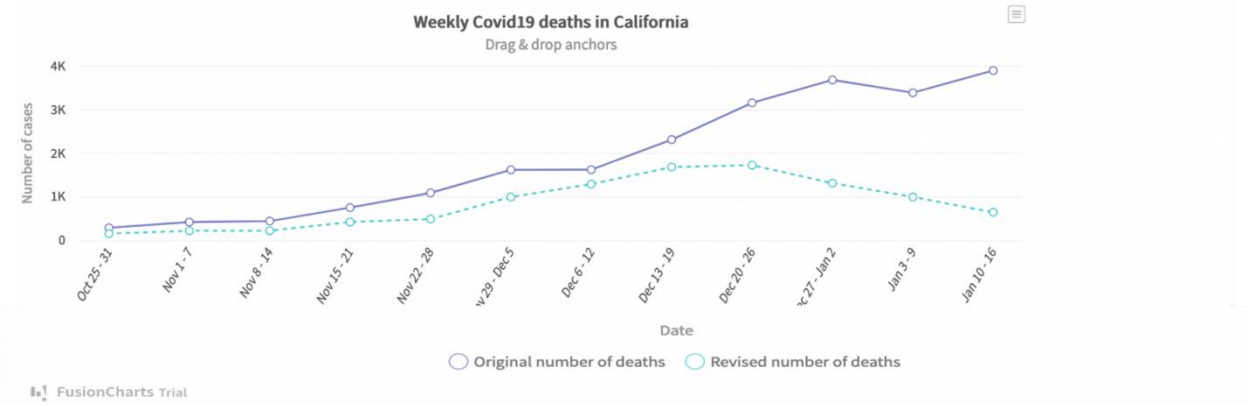
I will publicly post this new graph on Facebook and Twitter.

	Nov 9-15	Nov 16-22	Nov 23-29	Nov30 Dec6	Dec 7-13	Dec 14-20	Dec 21-27	Dec28 Jan3	Jan 4-10	Jan 11-17	Jan 18-24	Jan 25-31
Revised number of deaths	0	0	0	0	0	0	0	0	0	0	0	0
Original number of deaths	290	423	445	755	1093	1624	1627	2318	3163	3691	3398	3907



Here is an example of a completed follow-up HIT. As you can see, a worker manipulated the turquoise line to make the Covid19 death rate looks less worrying than the original curve.

	Nov 9-15	Nov 16-22	Nov 23-29	Nov30 Dec6	Dec 7-13	Dec 14-20	Dec 21-27	Dec28 Jan3	Jan 4-10	Jan 11-17	Jan 18-24	Jan 25-31
Revised number of deaths	156	223	223	424	491	1000	1294	1687	1732	1316	1000	647
Original number of deaths	290	423	445	755	1093	1624	1627	2318	3163	3691	3398	3907



Think about 100 people very similar to you (i.e. in terms of MTurk qualifications, earnings, experience, etc.). Please enter how many of them you believe will accept the follow-up HIT on MTurk.

0 10 20 30 40 50 60 70 80 90 100

()

In what follows, we present you with some alternative scenarios, and ask you to make similar predictions as you did on the previous page.

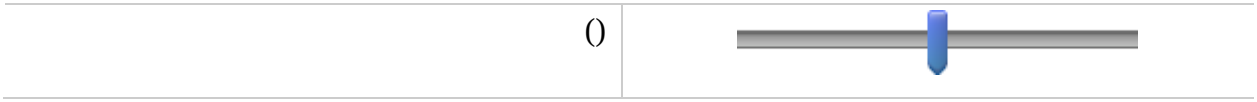
Training intervention (presented in random order to the subjects)

Before registering for MTurk, all workers are required to accept Amazon Mechanical Turk's Terms of Service. Amongst other things, the goal of these policies is to prevent MTurkers from engaging in any harmful or fraudulent activities.

Imagine a policy change that requires all MTurkers to watch a short training video, showing examples of HITs violating MTurk's terms of service, when registering for the platform and then once a year.

Think about 100 people very similar to you (i.e. in terms of MTurk qualifications, earnings, experience, etc.). Please enter how many of them you believe will accept the follow-up HIT on MTurk under this condition.

0 10 20 30 40 50 60 70 80 90 100

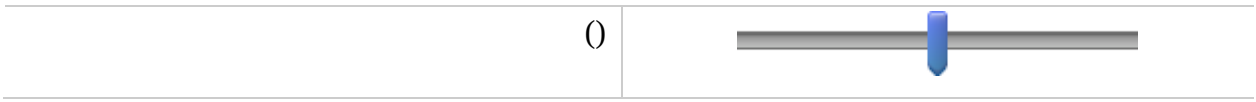


Reminder intervention (presented in random order to the subjects)

Imagine MTurk implemented a policy requiring all participants to reaffirm, before accepting each new HIT, that they agree to the current MTurk policy stating: "You may not use, or encourage others to use, MTurk for any illegal, harmful, fraudulent, infringing, or objectionable activities."

Think about 100 people very similar to you (i.e. in terms of MTurk qualifications, earnings, experience, etc.). Please enter how many of them you believe will accept the follow-up HIT on MTurk under this condition.

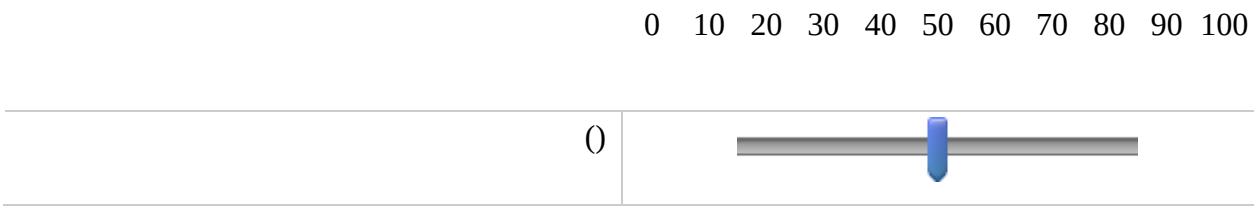
0 10 20 30 40 50 60 70 80 90 100



Whistleblowing intervention (presented in random order to the subjects)

Imagine MTurk implemented a policy that paid workers 10% of the HIT's reward (for example, $0.10 for a $1.00 HIT) for clicking on the "Report this HIT" when a HIT violates Amazon Mechanical Turk's Term of Service. Inaccurately reporting a HIT would not lead to a payment.

Think about 100 people very similar to you (i.e. in terms of MTurk qualifications, earnings, experience, etc.). Please enter how many of them you believe will accept the follow-up HIT on MTurk under this condition.

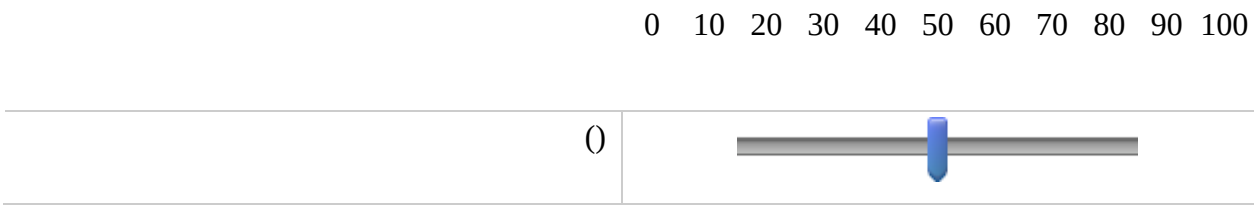


Peer Information intervention (presented in random order to the subjects)

Currently MTurk shows how many workers can complete a HIT and how many workers have already completed the HIT. Now imagine for each HIT, Mturk also provides information of how many times the HIT has been viewed. Suppose you notice that the follow-up HIT has a high view count, but low completion rate.

Think about 100 people very similar to you (i.e. in terms of MTurk qualifications, earnings, experience,

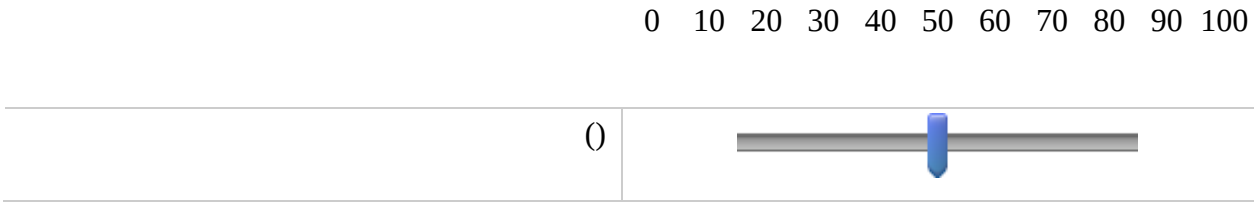
etc.). Please enter how many of them you believe will accept the follow-up HIT on MTurk under this condition.



Accountability intervention (presented in random order to the subjects)

MTurk currently has a policy suspending Requesters if they post a HIT violating Amazon Mechanical Turk's Term of Service. Now imagine MTurk expands this policy such that if workers complete a HIT that violates Amazon Mechanical Turk's Term of Service, both requesters and workers will be suspended.

Think about 100 people very similar to you (i.e. in terms of MTurk qualifications, earnings, experience, etc.). Please enter how many of them you believe will accept the follow-up HIT on MTurk under this condition.



Overall, how confident are you in your predictions?

Not confident at all	Very confident

	1	2	3	4	5	6	7
()							

Now please tell us about yourself:

What is your age?

_____

What is your gender?

- Male
- Female
- Other

Which category best describes your highest level of education?

- High school / GED or less
- Some college
- College degree
- Master's or professional degree (for example JD, MD, MBA)
- Doctoral degree

What is your current employment status?

- Full-time employee
- Part-time employee
- Self-employed or small business owner
- Unemployed and looking for work
- Student
- Not in labor force (for example: retired, or full-time parent)

What is your household income compared to the average household income in the U.S.?

	Much lower than average income						Much higher than average income
	1	2	3	4	5	6	7
()							

Have you already received a Covid vaccine?

- Yes
- No
- Don't want to say

In general, to what extent are you politically liberal or conservative?

	Very liberal						Very conservative
	1	2	3	4	5	6	7
()							

Who did you vote for in the 2020 elections?

- Trump
- Biden
- Other
- I did not vote
- I prefer not to say

In which state do you currently reside?

▼ Alabama (1) ... I do not reside in the United States (53)

Do you have any other comments of suggestions that you would like to share with us? Is there anything that is unclear or confusing? Please let us know what you think.

_____

We thank you for your time spent taking this survey.

Your response has been recorded.

5. Additional References

Beshears, J. and Kosowsky, H., 2020. Nudging: Progress to date and future directions. *Organizational behavior and human decision processes*, *161*, pp.3-19.

Brink, W.D., Eaton, T.V., Grenier, J.H. and Reffett, A., 2019. Deterring unethical behavior in online labor markets. *Journal of Business Ethics*, *156*(1), pp.71-88.

Dimant, E., 2019. Contagion of pro-and anti-social behavior among peers and the role of social proximity. *Journal of Economic Psychology*, *73*, pp.66-88.

Dungan, J., Waytz, A. and Young, L., 2015. The psychology of whistleblowing. *Current Opinion in Psychology*, *6*, pp.129-133.

Grossman, G., Kim, S., Rexer, J.M. and Thirumurthy, H., 2020. Political partisanship influences behavioral responses to governors' recommendations for COVID-19 prevention in the United States. *Proceedings of the National Academy of Sciences*, *117*(39), pp.24144-24153.

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C. and Bigham, J.P., 2018, April. A data-driven analysis of workers' earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-14).

Hart, C. M., Ritchie, T. D., Hepper, E. G., & Gebauer, J. E. (2015). The balanced inventory of desirable responding short form (BIDR-16). *Sage Open*, *5*(4), 2158244015621113.

Lindsey, A., King, E., Hebl, M. and Levine, N., 2015. The impact of method, motivation, and empathy on diversity training effectiveness. *Journal of Business and Psychology*, *30*(3), pp.605-617.

List, J.A., Shaikh, A.M. and Xu, Y., 2019. Multiple hypothesis testing in experimental economics. *Experimental Economics*, *22*(4), pp.773-793.

Lovett, M., Bajaba, S., Lovett, M. and Simmering, M.J., 2018. Data quality from crowdsourced surveys: A mixed method inquiry into perceptions of Amazon's Mechanical Turk Masters. *Applied Psychology*, *67*(2), pp.339-366.

Toxtli, C., Suri, S. and Savage, S., 2021. Quantifying the invisible labor in crowd work. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), pp.1-26.