The Effects of Autonomy and Task meaning in Algorithmic Management of Crowdwork

Yuushi Toyoda
Fujitsu Laboratories
Kanagawa, Japan
toyoda.yuushi@fujitsu.com

Gale Lucas
USC Institute for Creative Technologies
Playa Vista, CA, USA
lucas@ict.usc.edu

Jonathan Gratch
USC Institute for Creative Technologies
Playa Vista, CA, USA
gratch@ict.usc.edu

ABSTRACT

With the tremendous development of AI technologies, people will increasingly encounter software algorithms that supervise their work. Algorithmic management is the term for AI that performs the functions traditionally reserved for human managers (hiring, firing, providing evaluative feedback, and setting compensation). Although such algorithms indisputably perform management functions, they are often framed as support tools that facilitate worker autonomy. Perceptions of autonomy can enhance productivity, especially when the work holds intrinsic meaning for workers. But crowdwork often seems meaningless. More problematically, the meaning of the work must sometimes be obscured due to reasons of security or experimental control (when the workers serve as subjects in a psychological experiment). In this paper, we conduct an online experiment (N=560) to investigate how autonomy-perceptions and the meaningfulness of work interact to shape crowdworker motivation. As predicted, we find that workers are motivated when their work has meaning and algorithmic management is framed in a way that makes worker autonomy salient. However, when work holds no meaning, we find productivity is enhanced when algorithms are framed in a way that makes algorithm control salient. We also find evidence that providing meaning to the work can introduce systematic biases in crowdworker responses that could undermine accuracy in certain contexts.

KEYWORDS

Worker motivation; Crowdwork; Feedback system; Self determination theory

ACM Reference Format:
Yuushi Toyoda, Gale Lucas, and Jonathan Gratch. 2020. The Effects of Autonomy and Task meaning in Algorithmic Management of Crowdwork. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 9 pages.

1 INTRODUCTION

In the 18th century, Henry David Thoreau wrote "men have become tools of their tools." While Thoreau meant this as a metaphor, 21st century AI tools are literally tasking, evaluating and compensating a growing number of human workers. The emerging field of "algorithmic management" explores how best to automate the functions of human managers [17, 38]. This includes traditional AI problems such as task allocation [8, 13], optimizing workflow [16] and evaluating success [28, 38]. But as human workers cannot (and should not) be seen simply as machines, AI research into algorithmic management must confront more human-centric questions. These questions include how algorithms can best motivate workers [37], how fairness be ensured [5, 27], and what are the ethical and societal implications of these techniques [19]. In this article, we focus on these human-centric questions. Specifically, we experimentally examine alternative techniques to maintain crowdworker motivation when their work is assigned, evaluated, and compensated by an algorithm. Enhancing user motivation and performance through human-agent interaction is an important challenge, not only for algorithmic management, but in a variety of AI disciplines including educational technology, persuasive technology for health, computer games, and personal productivity monitoring, and crowdsourcing. Thus, findings have potential implications for the design of a wide range of algorithmic techniques.

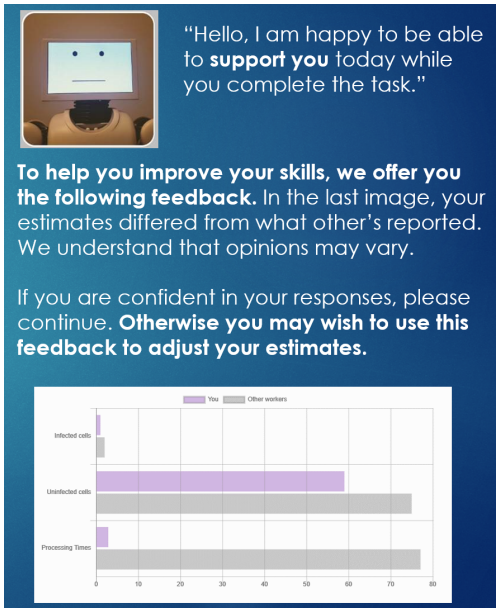

Figure 1: Our support agent which supervises workers during the task

There are choices for how to present the algorithmic manager to the workers. Prior research [22, 31] investigated how algorithmic management behind companies like Uber manages workers and frames the software to workers. Although traditional companies

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

use power hierarchies as a way to exert control to motivate worker productivity, algorithmic management is often framed in ways that obscure traditional power hierarchies (or more colorfully, Uber and Lyft "dissolve their agency and authority into an indifferent, automated algorithm" [35]). For example, the rhetoric around Uber's software is a support tool that promotes work autonomy and helps independent contractors do their business [31]. The promotion of entrepreneurship and autonomy through their app-based platform has been proven broadly successful. Yet, companies like Uber exact significant control over their workers by utilizing their algorithmic management functions such as constant tracking, constant evaluation of performance, and automatic implementation of penalty [22].

Autonomy can be a powerful motivator, but only if a worker is intrinsically motivated to perform the job (otherwise, they would use that autonomy to achieve other goals). One common way to enhance intrinsic motivation is to provide work meaning, for example, by providing the rationale underlying the work, or by highlighting its societal benefits [2]. These studies suggest that if workers find the intrinsic meaning of a task for them, workers are more engaged in the task and also produce higher quality output.

Although one might be tempted to conclude that one can always enhance crowdworker output by enhancing perceived autonomy and meaning, there are important circumstances where this may be unachievable or undesirable.

First, it is well known in the social sciences that providing a task meaning can bias the type of responses that workers provide. In what is known as a "response bias" [12, 24] or "demand characteristics" [23], participants in an experiment often seek out clues from the experimental instructions (or even experimenter nonverbal behavior) on how the experimenter wants them to respond. Often instructions intended to create meaning hold the potential for such bias. For example, if a worker is told to identify cancer cells in images to help cure cancer, it is possible they might be primed to find more false positives than a participant that was not told the meaning of the study. Given that much crowdwork is designed for such experimental purposes, this can be a real concern. Thus, there may be circumstances where one prefers to withhold meaning.

Second, some scholars have raised ethical arguments that algorithmic management must be framed in a way that makes its control functions explicit and transparent. For example, the uproar over Facebook's "emotional manipulation" study highlights that algorithms are framed in a way that makes their functions "invisible" and thus undermines a worker's ability to interrogate or negotiate these functions [11, 21]. Indeed some have argued the true motivation is to avoid giving gig workers full rights as employees [22, 31]. Therefore, there may be circumstances where one wishes to explicitly frame an algorithm as a supervisor. Our goal in this paper is not to discuss the ethics of framing algorithmic management as a supervisor or support, but to examine the practical impact of this framing on worker behavior.

In this work, our primary goal is to examine how crowdworker engagement is shaped by the meaningfulness of their task and the salience of the managerial power of algorithmic management (i.e., is the algorithm explicitly framed as a supervisor vs. a support tool). We examine this in the context of an online task where workers are overseen by an algorithm that exercises the traditional tools of a human supervisor (evaluates worker performance, provides corrective feedback when they make errors, and compensates them based on their performance).

A second goal is to seek evidence that crowdworker responses can indeed be biased when their task is provided meaning. To examine this, we adopt a task and a meaning manipulation that has already been discussed in the literature, but include new measures.

Finally, our work introduces a new, and we argue, purer measure of worker engagement with a task. Prior work has used some combination of worker output and accuracy to demonstrate that workers are more motivated by a task [2, 30]. But if we allow that accuracy may be systematically biased by how tasks are motivated, this highlights confound into the interpretation of our results.

2 RELATED WORK

Our research builds on several lines of existing work including theories of human motivation and human bias. We review these before outlining our hypotheses.

Worker Autonomy. Research on how to best motivate crowdworkers has much of its roots in the Self-Determination Theory (SDT) [32]. SDT is a theory of motivation that argues people are most productive and satisfied when initiating an activity for its own sake because it is interesting and satisfying in itself (intrinsic motivation), as opposed to doing an activity to obtain an external goal, such as money or obedience to authority (extrinsic motivation). Intrinsic motivation, according to SDT, is enhanced by promoting several basic physiological needs, including a person's need for autonomy. Autonomy is undermined by external coercive power (such as a human supervisor) or financial incentives [7]. It can even be influenced by factors as subtle as the use of autonomy-supportive language ("you can" or "you might") versus controlling language ("you must" or "you had better") [36].

SDT has been applied to various technological systems including educational technology [10, 29], persuasive technology for health [3, 4], computer games [25], personal productivity monitoring [15, 39], and crowdsourcing [14]. For example, Vansteenkiste et al. [36] showed that adolescent students were more engaged and learned better with a tutoring system that used autonomy-supportive language compared with controlling language. Consistent with SDT's arguments that external coercion can undermine intrinsic motivation, Mason and Watts found that increasing worker pay failed to increase the quality of crowdwork [20].

While the benefits of supporting worker autonomy are well-studied in human-human interactions, the connections between SDT and crowdwork is less explored, and much of that research has focused on factors other than autonomy. Further, we are unaware of any work that explicitly manipulates the salience of algorithmic management's coercive power (i.e., is it framed as a support tool versus a supervisor). The present study attempts to fill this gap.

Meaningfulness of Work. SDT is typically studied in contexts where the effort holds some value for the individual (e.g., learning a new skill or losing weight). Yet much of crowdwork seems devoid of meaning (filling in surveys or labeling images). Providing workers more autonomy may not enhance motivation if the task itself is inherently demotivating. To address this, research has

explored how to enhance intrinsic motivation by providing a meaningful rationale for the work. For example, Chandler and Kapelner [2] examined how crowdworker effort changed when they were provided a rationale for their work. Workers in the meaningful treatment group were told that they were labeling tumor cells in order to advance medical research whereas workers in the zero-context control group were not told any purpose of the task. They found that crowdworkers that were provided meaning were more likely to participate and the quantity of output increased. Consistent with this finding, Rogstadius et al. [30] found that worker output quality improved by telling them that their output helped a non-profit organization dedicating to curing malaria. Consistent with SDT's arguments about external coercion, they also found increasing payment doesn't improve the worker's output quality.

Taken together, research on autonomy and meaningfulness of work suggests that worker output will be maximized when algorithmic management supports worker autonomy (e.g., is framed as a support tool and provides autonomy-supporting feedback) and work is provided a meaningful rationale. Yet research on SDT has rarely examined how best to motivate workers when the work holds no meaning. One study did provide some insight into this question. As a strategy to motivate people for boring tasks, Deci et al. [6] examined the effects of three motivational factors related to intrinsic motivation: (a) providing a meaningful rationale, (b) acknowledging the participant's perspective, and (c) conveying choice rather than control. The study showed that when an environment supports at least two of these factors, participants were more engaged compared to the environment that supports one (one-factor condition) or zero (zero-factor condition) factors. However, contrary to their expectations, participants in the zero-factor condition were more engaged in the boring task than the participants in the one-factor condition. This suggests that framing algorithmic management as a supervisor (i.e., making the controlling aspects of the algorithm salient) could be of benefit in meaningless tasks, however, Deci and colleagues noted "further work is required to determine if this is a replicable finding."

Response Bias. Providing work meaning could inadvertently shift the type of responses that workers provide. Within the social sciences, the concept of response bias (see also "demand characteristics" or the "observer expectancy effect") is the phenomena where participants in an experiment form an interpretation of the experiment's purpose and subconsciously change their behavior to fit that interpretation [23]. Perhaps the most famous example of this is "Clever Hans", a horse that could correctly answer complex math problems. It was eventually discovered that the horse was sensitive to the experimenter's nonverbal cues and failed to answer correctly if the experimenters themselves didn't know the correct answer. Response bias is the reason why experiments are "double blind", meaning neither the experimenter nor the subjects are aware of key aspects of the study. Manipulations of meaning could inadvertently introduce response bias when some responses are seen as more socially desirable in and of themselves or more beneficial to achieving a desirable goal (such as curing malaria). We are unaware of research that has explored how task meaning shapes crowdworkers response bias. The present study fills this gap and explores this question with actual scientific data from a real-world medical task.

3 HYPOTHESES AND STUDY DESIGN

Our primary goal is to examine how algorithmic management can best enhance worker engagement and the quality of their output. Prior research on SDT suggests that worker engagement can be maximized when algorithmic management is framed as an autonomy-supportive tool and a meaningful rationale is provided for the work. This leads to our first hypothesis:

- H1: When a meaningful rationale of task is provided, crowdworkers will be more engaged by performance feedback given in an autonomy-supportive way by an autonomous agent.

Yet, despite the motivational benefits of infusing work with meaning, the introduction of meaning has the potential to shape the quality of worker output through response bias. This leads to our second hypothesis that providing meaning might have unanticipated negative consequences:

- H2: When a meaningful rationale of task is provided, crowdworker performance will be biased.

Given that meaning might have to be withheld, either to avoid bias or protect trade secrets, it is important to understand the best way to shape autonomy perceptions when work holds little meaning. Prior research on SDT is less clear about the impact of autonomy in these contexts, but as one study above suggests that algorithms could enhance worker feedback by restricting worker autonomy, we make the following third hypothesis:

- H3: When any context of task is not provided, workers will be more engaged by performance feedback given in a controlling way by an autonomous agent.

We examine these hypotheses with a 2 (task instruction: meaningful context vs. no context) x 2 (feedback design: supervisor vs. support agent) between-subjects experiment.[1]

3.1 Participants

We conducted our study on Amazon's Mechanical Turk (MTurk) platform, which is a popular platform for crowdsourcing. We collected task responses from 560 U.S. based MTurk users. Five of them were excluded from analysis due to either incomplete data or failure to follow instructions, which left 555 participants (39% female) remaining for analysis. MTurkers average age was 34 years (ranged from 18 to 73). The experimental design and materials were reviewed and approved by our university's ethics board.

3.2 Malaria Parasite Task

To enhance the real-world relevance of our findings and to allow direct comparisons with prior research on algorithmic management, MTurkers were asked to perform an actual scientific task: counting human cells infected with the malaria parasite (Figure 2). This task, or very similar tasks, have been used in prior research on

[1] We had an additional factor intended to influence autonomy perceptions (frequency of feedback varying from every 5 images to every image). However we found that the factor did not significantly impact our dependent variables, thus we ignore this condition for the remainder of the paper.

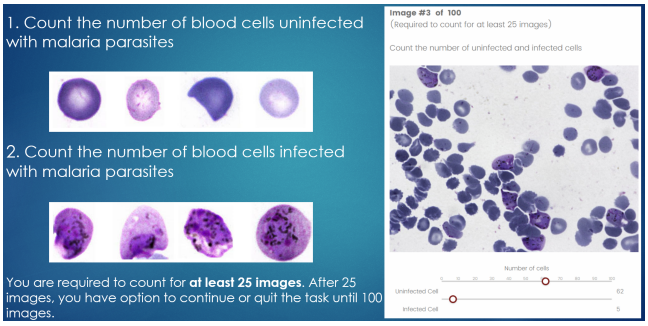

Figure 2: Counting task

crowdworker motivation [2, 30]. Although most prior research has used artificial data, to improve the real-world relevance of our conclusions we use actual medical images drawn from a publicly available corpus of medical images known as the Broad Bioimage Benchmark Collection [18]. We used the BBBC041v1 corpus of malaria images. Ground truth for these images was provided by a malaria researcher. One characteristic of such real data is it creates strong opportunities for workers to offer false positives. Specifically, there is a heavy imbalance towards uninfected cells (96% of all cells) including some images with no infected cells. Each image contained 64 uninfected cells on average (ranging from 42 - 91) and 2.64 infected cells on average (ranging from 0 to 8).

We asked workers to count the number of infected cells and uninfected cells in microscopy image sets. The image set included 100 images in total, with ground truth labels on the number of infected and uninfected cells. Workers were asked to count the cells in at least 25 images. After counting 25 images, the workers had an option to continue or quit the task. The workers could continue the task until 100 images. Instructions for the task, and many of our metrics, were adapted from two existing studies that manipulated the meaningfulness of work [2, 30]; we designed the experimental task to quantitatively measure the quality of work as well as quantity.

Following standard practice, workers received a fixed payment for their work (USD 5.00) plus a small incentive for each image they completed. The average worker took 30 minutes to complete the minimum request for 25 images. If they completed all 100 images they would receive USD 8.00.

3.3 Task Meaning

We manipulated the wording of the task instruction depending on the worker's condition, which was randomly assigned to meaningful context condition or no-context condition. For the workers in the meaningful context condition, we told the purpose of the task with images highlighting the importance of the task [18, 26] in the instruction. These instructions made it clear that it was important to identify the malaria parasites, thus creating the potential for response bias. On the other hand, we did not give any reason for their work in the no-context condition. The purpose of the task for the meaningful context condition was as follows:

> Thanks for participating in this task. Your job will be to help identify blood cells infected with malaria parasites in images and we appreciate your help. Malaria is a disease caused by Plasmodium parasites that remains a major threat in global health, affecting 200 million people and causing 1,000,000 deaths a year. 71% of all deaths are under age five. Besides biomedical research and political efforts, modern information technology is playing a key role in many attempts at fighting the disease. One of the barriers toward a successful mortality reduction has been inadequate malaria diagnosis in particular. To improve diagnosis, image analysis software and machine learning methods have been used to quantify parasite in microscopic slides. Accurate parasite counts are essential for malaria diagnosis.

As Figure 2 shows, we used several specific words related to malaria diagnosis in the task instructions for the meaningful context condition. On the other hand, for the no-context condition to remove any contexts of the task, we didn't use any words related to malaria diagnosis such as blood cells, malaria parasites, infected cells, and uninfected cells. Instead, we used generic words such as target objects and non-target objects in the task instructions as follows:

> This task requires you to count objects of different types. You will see a series of images containing two different types of objects. For each image, you are required to count the number of target objects and non-target objects. For each image,
> (1) Count the number of objects that appear most similar to the target class.
> (2) Count the number of objects that appear dissimilar to the target class.

3.4 Autonomy Perceptions

To manipulate perceptions of autonomy, the workers were randomly assigned to either the supervisor condition or support agent condition. Workers in the supervisor condition were told that they would be supervised by an artificially intelligent algorithm that was trained to execute all the tasks and functions which a human supervisor would normally perform. They were further told the AI supervisor would monitor their activity and provide evaluative feedback on their performance. Workers in the support agent condition were explained that they would be assisted by an AI support agent that was trained to execute all the tasks and functions which a human coach would normally perform. They were further told the AI would attend to their work and offer supportive feedback.

Both of the AI agents gave performance feedback to the workers based on how their answers compared with ground truth. Because we didn't want workers to feel their answers were irrelevant (which would be the case if the system already knew the ground truth), this feedback was described in comparison to how other workers performed on a particular image. For example, if workers estimate was quite different from ground truth (off by more than 3 infected cells or more than 20% off from the number of uninfected cells), workers were explained that their estimates differed from what

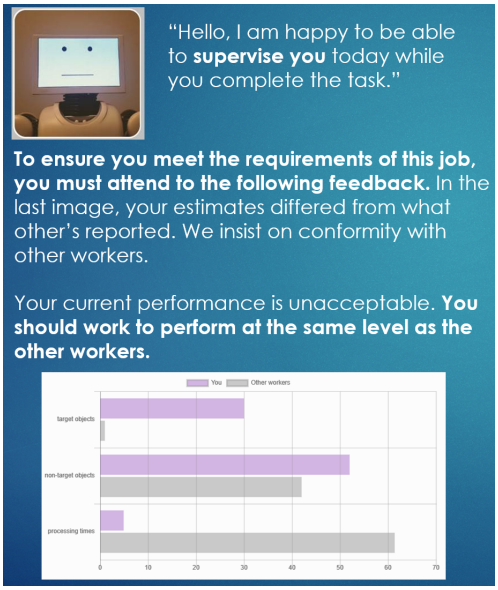

Figure 3: Performance feedback to the workers from the supervisor agent

other workers reported. If the estimate was close, the AI informed workers their estimates were in line with other workers. We also provided feedback on typical time-on-task (based on average times from our pilot study). Note that we used ground truth to judge cell counts, rather than the estimates from the pilot study, to better assess response bias.

We also manipulated the wording used in the feedback, following [36], to manipulate the salience of AI control. In the supervisor condition, the feedback used controlling language such as, "you should work to perform at the same level as the other workers," "you must attend to the following feedback" (Figure 3). In the support agent condition, the agent used autonomy-supportive phrases such as "you may wish to use this feedback to adjust your estimates," "to help you improve your skills, we offer you the following feedback" (Figure 1). Our agents are simple in appearance, without any gender, race, or other highly anthropomorphic traits that may trigger people's biases [1, 9, 33].

3.5 Dependent Measures

We measure both self-reported and behavioral measures of worker effort and motivation. We ask workers to self-report how much effort they expended and how motivated they were to perform the tasks (using 7-point Likert scales). More importantly, we collect several objective measures of worker output:

Time-on-task: We measured how much time on average a worker spent on each of the 25 initial images. Time-on-task can be a highly variable measure as MTurk users often multi-task and may be called away in the middle of a task. We excluded any images that took longer than 10 minutes to process.

Table 1: Correlation between task engagement and self-reported measures/processing times in the pilot study

	Spearman's rho	Sig.
Effort	.356	.000
Motivation	.184	.002
Processing Times	.654	.000

Task Engagement: Our primary goal is to examine how crowdworker motivation and output is shaped by perceptions of autonomy and meaning. Prior studies have looked at accuracy to quantify a worker's motivation. For example, Rogstadiusa and colleagues measure the number of incorrectly labeled cells normalized by the number of cells per image. However, as we hypothesize that accuracy can be influenced by meaning, accuracy may not be the best measure as it confounds worker motivation with response bias (a worker might be more motivated to produce quality work but have low accuracy due to their bias). To address this, we introduce a new measure, called task engagement that should be less influenced by response bias (we also report accuracy for comparison to past work). Task engagement, for a given MTurk user, was defined as the correlation between their reported number of cells (both infected and uninfected) and the true number across all 25 images:

$$TaskEngagement = c(x, y),$$

where

$$x = \{I_{est_1} + U_{est_1}, \ldots, I_{est_j} + U_{est_j}\},$$
$$y = \{I_{real_1} + U_{real_1}, \ldots, I_{real_j} + U_{real_j}\},$$

I is infected cells, U is uninfected cells, $real$ is ground truth of cells, est is estimated number by worker, j is image sequence number, $c(\cdot, \cdot)$ is a function of Spearman's rank correlation.

The rationale for this measure is it indicated how attentive the worker is to the characteristics of each image while being insensitive to scaling (e.g., if one worker always finds one extra cell than another worker, they will have the same task-engagement but differ in accuracy).

To validate this measure, we performed a pilot study to see if our measure of task engagement predicts self-reported motivation, self-reported effort, and actual time-on-task. We recruited 280 U.S. based MTurk users to count the number of infected cells and uninfected cells for 25 images. They were compensated USD 5.00 for their participation. Table 1 shows that task engagement is significantly and positively correlated with self-reported motivation, self-reported effort, and time-on-task. From this, we conclude it is a valid measure of motivation.

Accuracy: Because the set included some images without parasites, we could not use the metric for accuracy from previous work [30]. Besides, since the task was to count the number of cells (rather than labeling each cell), the standard accuracy measure for binary classification is not applicable. Therefore, we instead used the following metric to measure work quality:

$$Accuracy = 1 - \frac{|I_{est} - I_{real}| + |U_{est} - U_{real}|}{I_{real} + U_{real}},$$

where I is infected cells, U is uninfected cells, *real* is ground truth of cells, *est* is estimated number by worker.

Response Bias: To measure the response bias of the workers, we first coded each task response based on the reported infected cell count and the true number. Specifically, code 1 is given when the reported infected cell count is overestimated, code -1 is given when the reported infected cell count is underestimated, and code 0 is given when the reported infected cell count is the same as the true number. We then used the mean of the code in the 25 images as the response bias measure for each worker. Thus, a response bias measurement close to 1 indicates that the worker tends to overestimate, and close to -1 indicates that the worker tends to underestimate.

4 RESULTS

Before preceding to our main analysis of the behavior measures, we first performed two-way ANOVAs to assess the impact of autonomy and meaning on self-reported measures. We find no significant difference in self-reported motivation ($F(1, 551) = 2.464$, $p = .117$), or effort ($F(1, 551) = 1.021$, $p = .313$), though the means were quite high for both motivation (5.96 out of 7) and effort (6.36 out of 7), suggesting a possible ceiling effect.

Task Engagement: First, to further check the validity of our measure of motivation, we compared how well task engagement correlated with self-reported motivation, self-reported effort, and time-on-task. Replicating our pilot study, all of these correlations were significant and positive (Table 2). This reinforces confidence in this new measure.

We then performed a two-way ANOVA to assess the impact of perceived autonomy and task meaning on worker motivation (as indexed by task engagement). We found a significant cross-over interaction between these two factors on the level of task engagement ($F(1, 551) = 10.638$, $p = .001$). See figure 4. To break down this interaction, we investigated it further with a simple effects analysis. In the meaningful context condition, there is higher task engagement in the support agent condition ($\mu = .767$, $\sigma = .255$) than in the supervisor condition ($\mu = .701$, $\sigma = .313$). This marginally significant difference ($p = .067$) supports our hypothesis H1. In contrast, in the no-context condition, the results are reversed. There is significantly higher task engagement ($p = .005$) in the supervisor condition ($\mu = .783$, $\sigma = .242$) than in the support agent condition ($\mu = .691$, $\sigma = .301$). This is in line with our hypothesis H3.

Table 2: Correlation between task engagement and self-reported measures/processing times

	Spearman's rho	Sig.
Effort	.256	.000
Motivation	.157	.000
Processing Times	.606	.000

Further reinforcing these conclusions, we see in the support agent condition, there is significantly higher task engagement ($p = .031$) in the meaningful condition ($\mu = .767$, $\sigma = .255$) than in the no-context condition ($\mu = .691$, $\sigma = .301$). In the supervisor condition, on the other hand, there is higher task engagement in the no-context condition ($\mu = .783$, $\sigma = .242$) than in the meaningful context condition ($\mu = .701$, $\sigma = .313$). The difference is also statistically significant ($p = .014$).

In summary, these results suggest that workers were more engaged by autonomy-supportive feedback (than controlling feedback) when the purpose of the task was framed meaningfully, but workers were more engaged by controlling feedback (than autonomy-supportive feedback) when the task held no meaning.

Time-on-task: We performed a two-way ANOVA to assess the impact of perceived autonomy and task meaning on the time workers spent on the first 25 images. Further reinforcing the results of task engagement, we again find a significant cross-over interaction ($F(1, 551) = 4.161$, $p = .042$). As Figure 5 shows, in the support agent condition, workers in the meaningful context condition counted cells for a longer time ($\mu = 57.206$, $\sigma = 29.592$) than workers in the no-context condition ($\mu = 49.222$, $\sigma = 25.857$). The difference was statistically significant ($p = .013$). In addition, in the meaningful context condition, workers interacting with the support agent counted longer than workers with the supervisor agent ($\mu = 51.205$, $\sigma = 25.291$). The difference was marginally significant ($p = .064$). This again lends support to hypotheses H1 and H3.

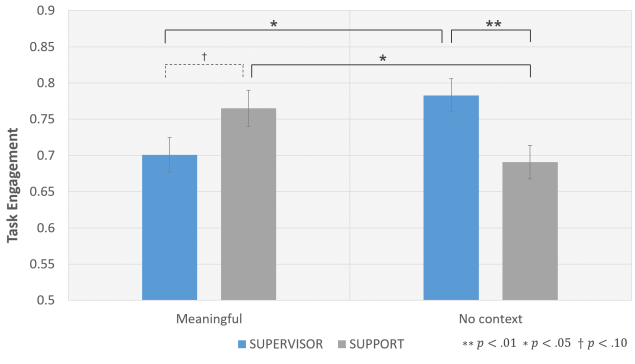

Figure 4: Task Engagement

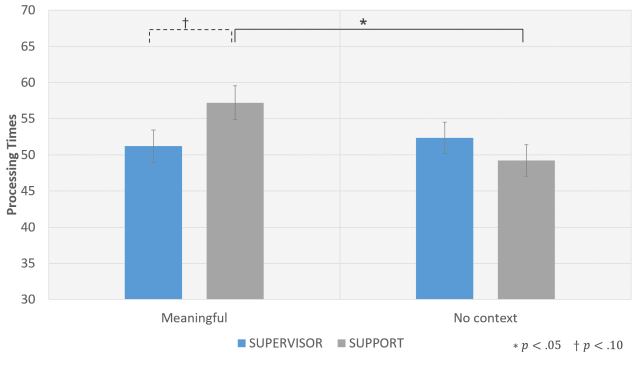

Figure 5: Time-on-task (seconds per image)

Table 3: Number of images counted by the conditions

	M_SV	M_SA	N_SV	N_SA
Participants	138	127	145	145
Counted images in 25 rounds	3450	3175	3625	3625
Counted images in 75 rounds	644	842	641	899
Counted images per worker	29.45	31.14	29.21	30.98

M_SV: Meaningful, Supervisor
M_SA: Meaningful, Support agent
N_SV: No context, Supervisor
N_SA: No context, Support agent

Work Quantity: Some crowdwork studies have looked at the quantity of output as a measure of motivation. Though findings on quantity have been mixed, we include for completeness. Table 3 shows the number of counted images by workers in each condition. Although workers with support agent counted images more than the workers with supervisor after the first 25 rounds, we didn't find a significant main effect for the feedback type ($F(1, 551) = 1.566$, $p = .221$), as well for the task instruction ($F(1, 551) = .051$, $p = .821$). There was no interaction between the two ($F(1, 551) = .004$, $p = .951$).

Accuracy: Next, we analyzed the accuracy of the task in the first 25 rounds. Although there were no main effects of the task instruction and the feedback design, we also found a significant interaction between the two factors ($F(1, 551) = 4.206$, $p = .041$). As Figure 6 shows, in the supervisor condition, workers in the no-context condition counted the blood cells more accurately ($\mu = .860$, $\sigma = .160$) than workers in the meaningful context condition ($\mu = .812$, $\sigma = .249$). The difference was statistically significant ($p = .027$). Also, in the meaningful context condition, workers interacting with the support agent counted more accurately ($\mu = .853$, $\sigma = .151$) than workers with the supervisor. The difference was marginally significant ($p = .063$). However, in the no-context condition, there was no significant difference, as the accuracy in the support agent conditions was similar to the supervisor condition. In the support agent condition, we found no significant difference between the no-context and the meaningful condition.

Table 4: Errors and accuracy by the conditions

		M_SV	M_SA	N_SV	N_SA
Uninfected cells (Absolute difference)	Mean	9.632	8.831	7.967	10.000
	SD	9.242	8.106	6.871	8.141
Infected cells (Absolute difference)	Mean	2.970	1.229	1.594	1.079
	SD	8.936	3.280	5.318	2.670
Accuracy	Mean	.812	.853	.860	.838
	SD	.249	.151	.160	.143

M_SV: Meaningful, Supervisor
M_SA: Meaningful, Support agent
N_SV: No context, Supervisor
N_SA: No context, Support agent

For a detailed analysis on the accuracy, table 4 shows the absolute difference between reported cell count and the ground truth for uninfected cells and infected cells. Regarding the absolute difference of uninfected cells, there was also a significant interaction between the task instruction and the feedback design ($F(1, 551) = 4.211$, $p = .041$). Simple effects analysis showed that at the no-context condition, workers interacting with support agent made more errors ($\mu = 10.000$, $\sigma = 8.141$) than workers with the supervisor agent ($\mu = 7.967$, $\sigma = 6.871$). The difference was statistically significant ($p = .033$). At the supervisor condition, workers in the meaningful condition ($\mu = 9.632$, $\sigma = 9.242$) made more errors than workers in the no-context condition. The difference was marginally statistically significant ($p = .085$). Regarding the absolute difference of infected cells, we found a significant main effect of the feedback design ($F(1, 551) = 5.572$, $p = .019$). Workers interacting with supervisor agent made more errors ($\mu = 2.135$, $\sigma = 6.879$) than workers with the support agent ($\mu = 1.336$, $\sigma = 4.208$).

Response Bias: Finally, we performed a two-way ANOVA to assess the impact of perceived autonomy and task meaning on response bias (see Figure 7). We found a significant main effect of meaningfulness ($F(1, 551) = 10.487$, $p = .001$) such that workers over-reported infected cells in the meaningful condition. There was no main effect of the autonomy manipulation on response bias ($F(1, 551) = .333$, $p = .564$), as well no interaction between the two factors

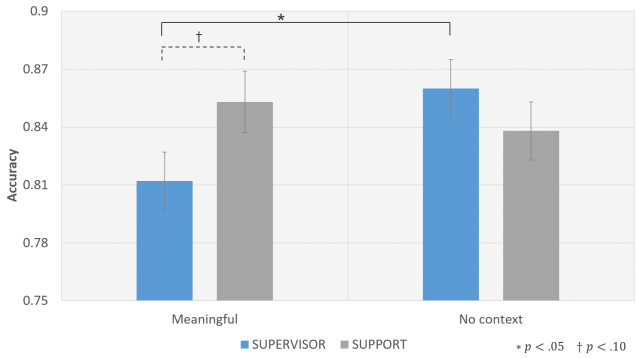

Figure 6: Accuracy

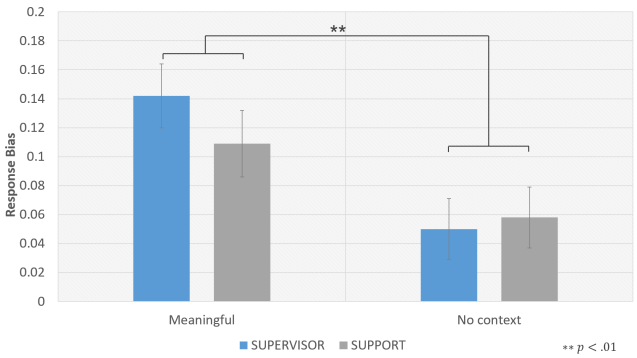

Figure 7: Response Bias

($F(1, 551) = .863$, $p = .353$). Workers in the no-context condition showed less response bias ($\mu = .054$, $\sigma = .257$) than workers in the meaningful condition ($\mu = .125$, $\sigma = .258$). This finding supports our second hypothesis (H2): workers who were provided the task meaning clearly overestimated the number of the infected cells compared with workers that were not told the meaning of the study.

5 DISCUSSION

The results provide strong confirmation of our hypotheses. Following self-determination theory, prior research [6, 34, 36] has hypothesized the importance of worker autonomy and meaning to enhance worker motivation (Hypothesis H1) and our study replicates prior research, thus lending further support to these claims. When work was provided a meaningful rationale, task engagement significantly increased when AI was framed as a support tool and used autonomy-supportive language. This also translated into greater time-on-task and a trend for greater accuracy. Interestingly, workers did not seem aware of these influences as there were no differences in their self-reported effort or motivation.

However, our study highlights the potential downsides of using meaning to enhance worker motivation. Following the literature on response bias and experimenter effects, we had hypothesized that providing meaning could shift workers responses towards a socially desirable response (Hypothesis H2). In our task, the desired goal was to help cure malaria by finding parasites in images of blood cells. Yet the number of parasites in this real scientific data was quite low (many slides contained no parasites at all), creating the opportunity for workers to provide false positives. And this is indeed what we found. Workers falsely identified more parasite cells when the meaning of the task was explained. This is an important qualification to studies that emphasize the importance of providing meaning [2, 30].

Given that there are situations where the meaning is withheld, it is important to clarify how best to motive workers in such situations. Prior research on self-determination theory is inclusive on this point but we had hypothesized that workers could be better motivated when AI was framed as a supervisor and used controlling language (Hypothesis H3). This hypothesis was supported. Workers were more engaged in meaningless tasks when the AI was framed as a supervisor (this framing did not impact time-on-task or accuracy). Again, workers were not aware of these influences (as assessed by their self-reported responses).

6 CONCLUSIONS

Our primary objective in this study was to examine how algorithmic management can best enhance worker engagement and the quality of their output. To achieve this goal, we conducted an online experiment with 560 crowdworker on Amazon Mechanical Turk, who were given either a meaningful context or no-context, and either autonomy-supportive feedback or controlling feedback on our malaria parasite task.

We found that there was a significant cross-over interaction between autonomy perceptions and meaningfulness of work. Specifically, our experimental results showed that workers are motivated when their work has meaning and algorithmic management is framed in a way that makes worker autonomy salient. While, when work holds no meaning, productivity is enhanced when algorithms are framed in a way that makes algorithm control salient. Furthermore, we indeed found that providing work meaning shapes crowdworkers response bias. Our work contributes to the growing body of literature in algorithmic management and human-AI interaction to better design the society where humans and AI work together.

ACKNOWLEDGMENTS

This work is supported in part by Fujitsu Laboratories, Fujitsu Laboratories of America, and the US Army. The content does not necessarily reflect the position or the policy of any Government, and no official endorsement should be inferred.

REFERENCES

[1] Jeremy N Bailenson, Jim Blascovich, Andrew C Beall, and Jack M Loomis. 2003. Interpersonal distance in immersive virtual environments. *Personality and Social Psychology Bulletin* 29, 7 (2003), 819–833.

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

[3] Eun Kyoung Choe, Bongshin Lee, Sean Munson, Wanda Pratt, and Julie A Kientz. 2013. Persuasive performance feedback: The effect of framing on self-efficacy. In *AMIA Annual Symposium Proceedings*, Vol. 2013. American Medical Informatics Association, 825.

[4] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al. 2008. Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1797–1806.

[5] Celso M de Melo, Stacy Marsella, and Jonathan Gratch. 2019. Human Cooperation When Acting Through Autonomous Machines. *Proceedings of the National Academy of Sciences* 116, 9 (2019), 3482–3487.

[6] Edward L Deci, Haleh Eghrari, Brian C Patrick, and Dean R Leone. 1994. Facilitating internalization: The self-determination theory perspective. *Journal of personality* 62, 1 (1994), 119–142.

[7] Edward L Deci and Richard M Ryan. 2010. Intrinsic motivation. *The corsini encyclopedia of psychology* (2010), 1–2.

[8] John P Dickerson, Karthik Abinav Sankararaman, Aravind Srinivasan, and Pan Xu. 2018. Assigning tasks to workers based on historical data: Online task assignment with two-sided arrivals. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 318–326.

[9] Ron Dotsch and Daniël HJ Wigboldus. 2008. Virtual prejudice. *Journal of experimental social psychology* 44, 4 (2008), 1194–1198.

[10] Sidney K D'Mello. 2016. Giving eyesight to the blind: Towards attention-aware AIED. *International Journal of Artificial Intelligence in Education* 26, 2 (2016), 645–659.

[11] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 153–162.

[12] Adrian Furnham. 1986. Response bias, social desirability and dissimulation. *Personality and individual differences* 7, 3 (1986), 385–400.

[13] Ece Kamar, Severin Hacker, and Eric Horvitz. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 467–474.

[14] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk.. In *AMCIS*, Vol. 11. Detroit, Michigan, USA, 1–11.

[15] Young-Ho Kim, Jae Ho Jeon, Eun Kyoung Choe, Bongshin Lee, KwonHyun Kim, and Jinwook Seo. 2016. TimeAware: Leveraging framing effects to enhance personal productivity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 272–283.

[16] Aniket Kittur, Susheel Khamkar, Paul André, and Robert Kraut. 2012. CrowdWeaver: visually managing complex crowd work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 1033–1036.

[17] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1603–1612.

[18] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. 2012. Annotated high-throughput microscopy image sets for validation. *Nature methods* 9, 7 (2012), 637–637.

[19] Caitlin Lustig, Katie Pine, Bonnie Nardi, Lilly Irani, Min Kyung Lee, Dawn Nafus, and Christian Sandvig. 2016. Algorithmic authority: the ethics, politics, and economics of algorithms that interpret, decide, and manage. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1057–1062.

[20] Winter Mason and Duncan J Watts. 2009. Financial incentives and the performance of crowds. In *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, 77–85.

[21] Robinson Meyer. 2014. Everything We Know About Facebook's Secret Mood Manipulation Experiment. https://www.theatlantic.com/. (2014).

[22] Mareike Mohlmann and Lior Zalmanson. 2017. Hands on the wheel: Navigating algorithmic management and uber drivers' autonomy. *In Proceedings of the 38th International Conference on Information Systems (ICIS)* (2017).

[23] Martin T Orne. 1962. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American psychologist* 17, 11 (1962), 776.

[24] Delroy L Paulhus. 1991. Measurement and control of response bias. (1991).

[25] Eli Pincus, Su Lei, Gale Lucas, Emmanuel Johnson, Michael Tsang, Jonathan Gratch, and David Traum. 2018. The importance of regulatory fit & early success in a human-machine game. In *Proceedings of the Technology, Mind, and Society*. ACM, 31.

[26] Mahdieh Poostchi, Kamolrat Silamut, Richard J Maude, Stefan Jaeger, and George Thoma. 2018. Image analysis and machine learning for detecting malaria. *Translational Research* 194 (2018), 36–55.

[27] Chenxi Qiu, Anna Squicciarini, and Benjamin Hanrahan. 2019. Incentivizing Distributive Fairness for Crowdsourcing Workers. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 404–412.

[28] Chenxi Qiu, Anna Squicciarini, Dev Rishi Khare, Barbara Carminati, and James Caverlee. 2018. CrowdEval: A Cost-Efficient Strategy to Evaluate Crowdsourced Worker's Reliability. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1486–1494.

[29] Jennifer Robison, Scott McQuiggan, and James Lester. 2009. Evaluating the consequences of affective feedback in intelligent tutoring systems. In *2009 3rd international conference on affective computing and intelligent interaction and workshops*. IEEE, 1–6.

[30] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *Fifth International AAAI Conference on Weblogs and Social Media*.

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

[32] Richard M Ryan and Edward L Deci. 2000. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology* 25, 1 (2000), 54–67.

[33] Mikey Siegel, Cynthia Breazeal, and Michael I Norton. 2009. Persuasive robotics: The influence of robot gender on human behavior. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2563–2568.

[34] Marlene N Silva, Paulo N Vieira, Sílvia R Coutinho, Cláudia S Minderico, Margarida G Matos, Luís B Sardinha, and Pedro J Teixeira. 2010. Using self-determination theory to promote physical activity and weight control: a randomized controlled trial in women. *Journal of behavioral medicine* 33, 2 (2010), 110–122.

[35] Julia Tomassetti. 2016. Does Uber Redefine the Firm: The Postindustrial Corporation and Advanced Information Technology. *Hofstra Lab. & Emp. LJ* 34 (2016), 1.

[36] Maarten Vansteenkiste, Joke Simons, Willy Lens, Kennon M Sheldon, and Edward L Deci. 2004. Motivating learning, performance, and persistence: the synergistic effects of intrinsic goal contents and autonomy-supportive contexts. *Journal of personality and social psychology* 87, 2 (2004), 246.

[37] Ming Yin and Yiling Chen. 2015. Bonus or not? learn to reward in crowdsourcing. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[38] Han Yu, Chunyan Miao, Yiqiang Chen, Simon Fauvel, Xiaoming Li, and Victor R Lesser. 2017. Algorithmic management for improving collective productivity in crowdsourcing. *Scientific reports* 7, 1 (2017), 12541.

[39] Nan Zhao, Asaph Azaria, and Joseph A Paradiso. 2017. Mediated atmospheres: A multimodal mediated work environment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 31.