Online Appendix for Improving Human Sequential Decision Making with Reinforcement Learning

Appendix C: Additional Details on Experimental Results

C.1. Pilot Studies

We ran small-scale pilot studies in late 2019 and early 2020 both online via AMT and in person at our institution's behavioral lab. Following best practices, the main objectives of these pilots were to finalize the design of the game and ensure feasibility, but not to estimate treatment effects. For example, we investigated how long it would take the participant to finish each round of the game, calibrated the dishes and their cooking tasks/sequences, figured out how to portray different virtual kitchen workers, and improved the user interface of the game. We also experimented with different framing and presentation of tips to get a preliminary understanding of how participants would notice and respond to the tips. Once we identified our final design of the game, we pre-registered our main study and started collecting data in Summer 2020.

C.2. Participant Demographics and Performances

Table C.1 exhibits the number of participants across our studies, and participants' demographic and gameplay information across these studies. *Attempted N* refers to the number of participants who started and attempted the game, while *Completed N* indicates the number of participants who successfully completed the game and are included in our analyses. The criteria for exclusion, as outlined in our pre-registration, include failing to follow instructions in the game, failing to complete any of the rounds within 45 steps, failing our attention check, providing an invalid AMT Worker ID or confirmation code generated at the completion of the game, or participating in more than one study. Importantly, we always ran all arms concurrently for each study, so that any variations in the participant pool affected all arms similarly.

Participants playing the disrupted configuration generally took longer to complete the game and found it to be more difficult, compared to the normal configuration. Tables C.2 and C.3 show the average performance in each round across phases and treatment conditions for normal and disrupted configurations, respectively.

	Phase I: Normal	Phase II: Normal	Phase I: Disrupted	Phase II: Disrupted
Attempted N	198	2,025	200	1,965
Completed N	183	1,317	172	1,011
Mean age [range]	35 [18, 76]	33 [18, 74]	34 [19, 76]	35 [16, 84]
Female	57%	51%	62%	60%
$\geq$ 2-year degree	73%	68%	78%	70%
Median duration	19 minutes	21 minutes	28 minutes	27 minutes
Found the game difficult	61%	50%	71%	65%
Never played similar games	45%	44%	47%	44%

Table C.1 Participants' demographic and gameplay information.

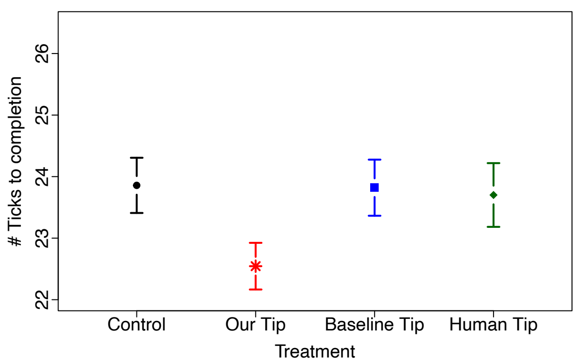
Figure C.1 illustrates participant performance in Phase II of our primary experiment in both the normal and disrupted configurations (same information as Figure 4 but in terms of raw ticks) but in the number of

	Phase I	Phase II: Control	Phase II: Algorithm	Phase II: Baseline	Phase II: Human
Round 1	25.73	26.03	25.04	26.01	26.16
Round 2	25.02	24.46	23.29	24.71	25.06
Round 3	23.74	23.86	22.99	24.04	24.06

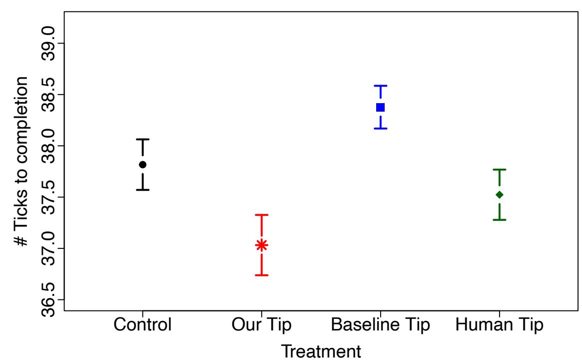
Table C.2 Average performance by treatment condition and round (normal configuration).

	Phase I	Phase II: Control	Phase II: Algorithm	Phase II: Baseline	Phase II: Human
Round 1	24.35	24.26	24.18	24.77	24.64
Round 2	22.87	22.38	22.95	23.08	22.69
Round 3	38.75	38.73	38.19	38.74	38.26
Round 4	38.39	38.21	37.77	38.38	37.85
Round 5	38.25	38.17	37.25	38.42	37.62
Round 6	37.96	37.82	37.14	38.37	37.62

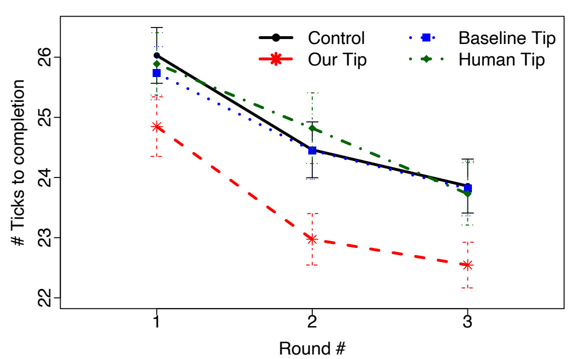
Table C.3 Average performance by treatment condition and round (disrupted configuration).



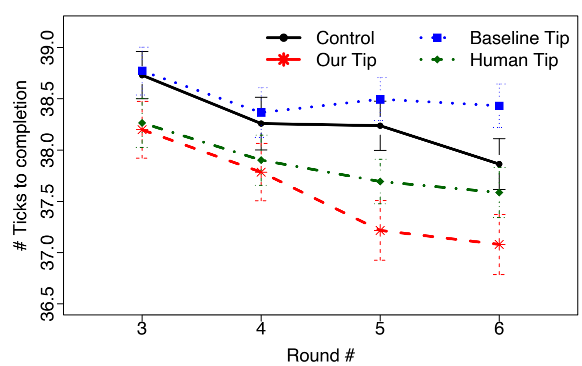
(a) Final Round Performance (Normal)



(b) Final Round Performance (Disrupted)

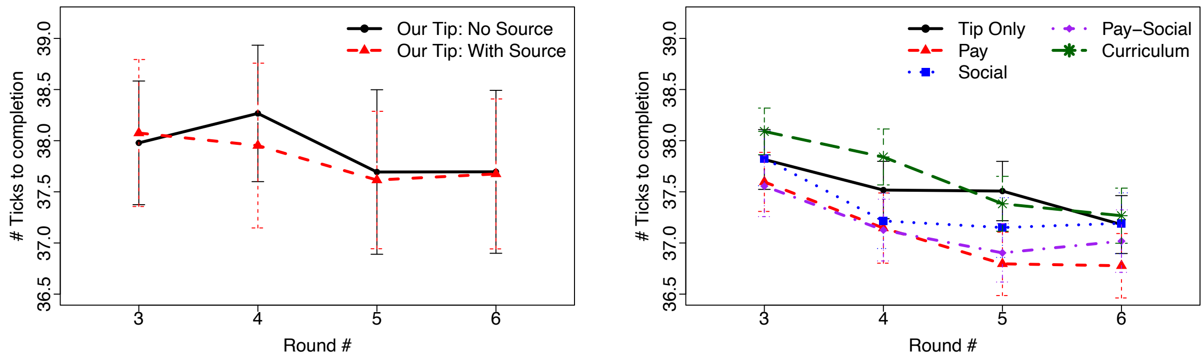


(c) Performance over Time (Normal)



(d) Performance over Time (Disrupted)

Figure C.1 Phase II Participant Performance. The subfigures depict various views of participant performance across conditions in the normal (left) and disrupted (right) configurations. The top row shows performance in the last round of the configuration, the bottom row shows how participant performance improves over time.



(a) Source of Tip (b) Interventions for Compliance

Figure C.2 Phase II Participant Performance for Follow-Up Studies. The subfigures depict performance over time across conditions in the Source of Tip (left) and the Interventions for Compliance (right) studies.

ticks instead of the relative percentage of ticks above optimal. Analogously, Figure C.2 illustrates participant performance in terms of raw ticks for our pilot experiment on revealing source information (Appendix C.6) and follow-up experiment on improving compliance (Section 5.4).

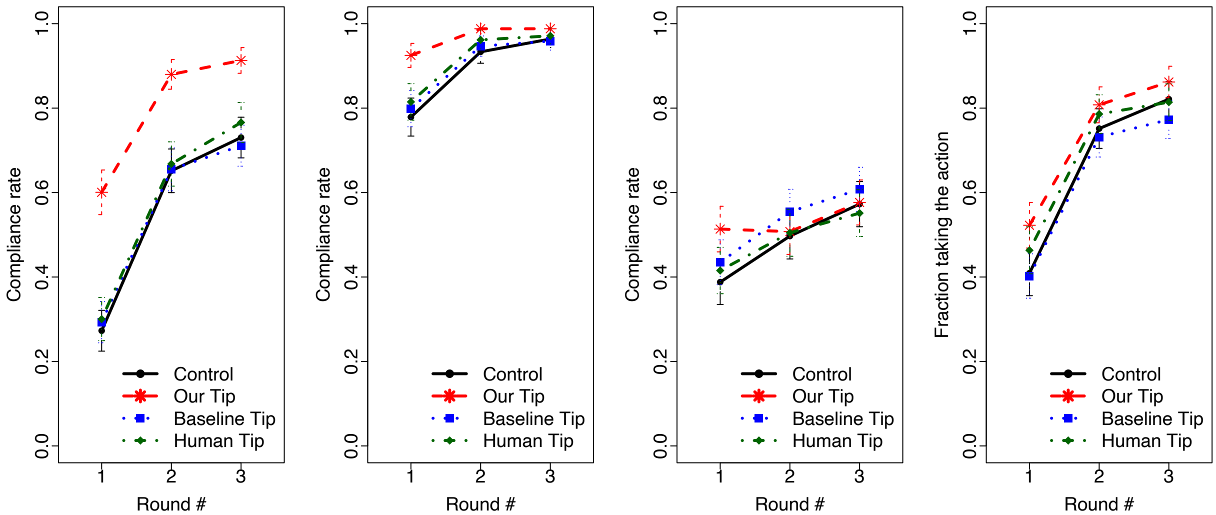
C.3. Learning Beyond Tips under The Normal Configuration

Compared to the disrupted configuration, the normal configuration is much easier. We find that participants across all conditions cross-comply with all other tips: not to assign plating to the chef (Figure C.3a), strategically leave some virtual workers idle (Figure C.3b), and let the chef chop only once (Figure C.3c). These tips are all consistent with the optimal policy, suggesting that participants generally learn over time to improve their performance regardless of the condition. Interestingly, participants in the algorithm condition have similar or higher cross-compliance compared to the other conditions. This result suggests that our tip is the most effective as the information it encompasses the information conveyed by the other tips. At a high level, the optimal policy for the fully-staffed scenario has the chef cook most of the dishes, has the server plate most of the dishes, and never assigns the chef to plate or the server to cook. We observe that participants generally recover these optimal strategies as they gain more experience with the game. For instance, the fraction of participants in each arm that never assign cooking to the server in each round, as if they were following the tip "Server shouldn't cook", increases over time and within each round the fractions are not statistically different among the arms (see Figure C.3d). This result suggests that participants can uncover this unshown rule by themselves across all conditions.

C.4. Robustness Checks for Tip Construction

We now examine the robustness of our results to several design choices.

Robustness to subpopulation used to construct human tip. In our experiment, we desginated the human tip as the highest-voted tip among *all* participants from Phase I. However, one may hypothesize that peer feedback is more effective if it is inferred from *top* performers. To this end, we examine what tips we would have derived if we had restricted to the highest-voted suggestion by the (absolute top, top 5%, top 10%, and top 25%) of Phase I performers. We define performance "Top X% performers" refers to the Phase I



(a) Our Tip: "Chef Shouldn't Plate"
(b) Human Tip: "Leave Some Workers Idle"
(c) Baseline Tip: "Chef Chops Once"
(d) Unshown Tip: "Server Shouldn't Cook"

Figure C.3 Learning beyond Tip (Normal Configuration). Panels (a)-(c) show the rate at which participants in each condition cross-comply with each offered tip over time in the normal configuration. Panel (d) shows analogous results for a rule that is part of the optimal policy but was not shown as a tip in any condition.

participants who had the highest performance in Round 6 (final round). The human tip remained the *same* across all of these subgroups in both configurations; in other words, focusing on best performers would not change our results, since we would still identify the same human tips (see Table C.4).

Threshold for elimination	Normal configuration	Disrupted configuration
Everyone (Original)	Leave some idle	Server cooks once
Best performers	Leave some idle	Server cooks once
Top 5% performers	Leave some idle	Server cooks once
Top 10% performers	Leave some idle	Server cooks once
Top 25% performers	Leave some idle	Server cooks once

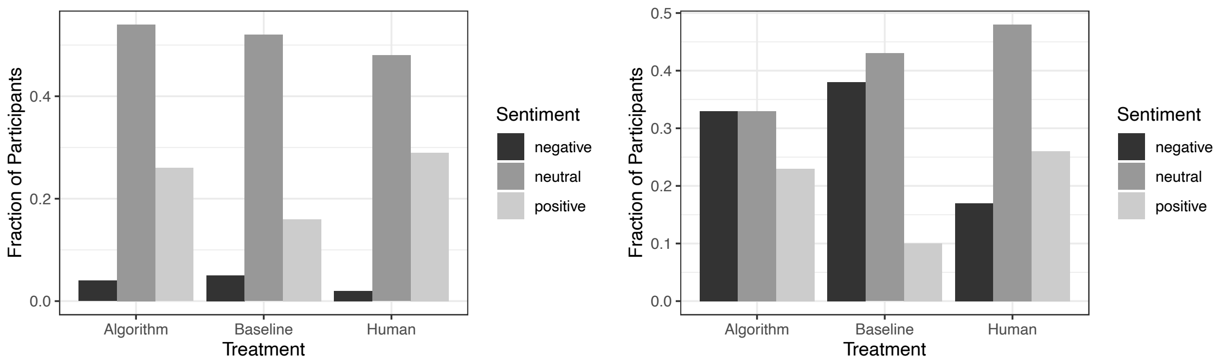
Table C.4 Top tips for the "Human" condition based on various subpopulations

Robustness of our algorithm's tip to varying quantiles of human trace data. As described in Appendix A.4, for computing our algorithm's tip, we trained on the bottom 25% of participants since the goal of our paper is to help improve the performance of workers who are weakest at the given problem. Indeed, our expected improvement is much higher for the bottom 25% (3.6 tips faster for normal, 4.4 ticks faster for disrupted) than for everyone (2.1 ticks faster for normal, 1.8 ticks faster for disrupted), demonstrating that our tip is expected to be most effective for the bottom quartile of participants. To analyze the robustness of our approach, we considered two alternatives: using the bottom 50% or using everyone. As shown in Table C.5, our algorithm produces the same tips using these alternative strategies.

Robustness of our algorithm's tip to elimination threshold. As described in Appendix A.4, for computing our algorithm's tip, we used a post-processing step where we pruned tips that that disagree with the optimal policy more than 50% of the time. To evaluate robustness, we now consider constructing our algorithm's tip

Criteria for tip selection	Normal configuration	Disrupted configuration
Bottom 25% (Original)	Chef should never plate	Server cooks twice
Bottom 50%	Chef should never plate	Server cooks twice
Everyone	Chef should never plate	Server cooks twice

Table C.5 Top tips by our algorithm based on varying quantiles of Phase I human trace data



(a) Normal Configuration (b) Disrupted Configuration

Figure C.4 Participant sentiment on the provided tips in post-game survey.

based on alternative thresholds of 30% and 70%. Results are shown in Table C.6; as can be seen, the tip is robust to the choice of this threshold.

Threshold for elimination	Normal configuration	Disrupted configuration
30%	Leave some idle	Server cooks once
50% (Original)	Leave some idle	Server cooks once
70%	Leave some idle	Server cooks once

Table C.6 Top tips for the "Human" condition by various elimination criteria

C.5. Sentiment Analysis of Participants' Qualitative Responses Using Machine Learning

To better understand how participants perceived our tips, we conducted a sentiment analysis for the qualitative responses to the post-study survey. Our analysis is based on manual coding (by an independent human who is not part of the research team) of each participant's sentiment towards the tip they received—positive, neutral, or negative—based on their post-game survey responses to the question *"What did you think about the tip for these last [three/four] rounds and how did you incorporate it in your strategy?"*. Figure C.4 shows the breakdown of responses in each condition and configuration.[13]

First, we observe that for both configurations, generally more participants in the human condition responded positively and fewer responded negatively to peer-derived tips compared to the other conditions.[14]

[13] We excluded unrelated/uninformative participant responses, so fractions per condition may not add to 1.

[14] Using Pearson's χ^2 tests: For the normal configuration, more participants in the human condition responded positively to the tips compared to the algorithm ($\chi^2(1) = 0.78,\ p = 0.19$) and the baseline ($\chi^2(1) = 16.259,\ p < 10^{-4}$) conditions, while fewer in the human condition responded negatively compared to the algorithm ($\chi^2(1) = 2.19,\ p = 0.07$) and the baseline ($\chi^2(1) = 4.59,\ p = 0.02$) conditions. For the disrupted configuration, more participants in the human condition responded positively to the tips compared to the algorithm ($\chi^2(1) = 0.61,\ p = 0.2$) and the baseline ($\chi^2(1) = 18.91,\ p < 10^{-4}$) conditions, while fewer in the human condition responded negatively compared to the algorithm ($\chi^2(1) = 17.63,\ p < 10^{-4}$) and the baseline ($\chi^2(1) = 28.84,\ p < 10^{-4}$) conditions.

In other words, human participants selected tips that would likely be accepted by other humans; these tips offer natural strategies that match human intuition, e.g., we observe that they are often adopted even in the control group. On the other hand, the algorithmic and baseline tips are necessarily part of the optimal (rather than human) policy, and can therefore be counter-intuitive; this is especially apparent in the disrupted configuration, where the algorithm and baseline tips both received substantially more negative feedback ($p < 10^{-4}$ for both comparisons), and therefore lower compliance rates, as observed in Figure 5b. However, performance and compliance improved over time, implying that it took participants time and effort to correctly incorporate these tips into their workflow and execute an optimal strategy. This is supported by selected excerpts from participant comments presented in Table C.7.

Disrupted Configuration	Our Algorithm's Tip *"Server should cook twice"*	Human Tip *"Server should cook once"*
Positive	• "It was very helpful. It made me focus on making sure the server cooked more even if that was not his obvious strength." • "I ignored the tip at first, but later I used the tip and it helped me complete the tasks quickly." • "At first I didn't follow it because it seemed counter intuitive since they're slow. But then I had trouble, so I tried it and came out ahead." • "I did not listen to it at first because I didn't believe that it would actually help but it did." • "The tip was helpful. Without it, I think I would have tried to complete the task without the Server cooking, which would have left someone idle for a long time."	• "It seemed pretty much essential to have server cook once." • "I thought it was smart and I used it exclusively." • "It was accurate, and I implemented the tip." • "I felt that tip was valid, as the server primarily is useful plating/ chopping. I only had him cook once." • "It helped because she could cook one burger but any more than that and your ticks would be too high."
Negative	• "I think it was a bad tip. I couldn't figure out how to incorporate it successfully." • "Seemed counterintuitive." • "It did not help me. I did not use it for round 1, I used it for round 2 and it made me do worse, so round 3 I tried it again and was still unable to do well, so the last round I ignored the tip." • "I don't think it helped. I thought having the sous chef cook 3 times would take too long and the point at which I tried it, I decided last minute to have the server cook twice. So I don't think it told me anything useful." • "It was not needed since the server took so much longer to cook."	• "It was not helpful, because it does not specify when the server should cook." • "I used the tip but I don't think it was helpful. The server took long to cook." • "I don't agree with this tip." • "It was not terribly helpful. I tried to incorporate but it did not seem to help" • "It stunk honestly. The server takes forever to cook."

Table C.7 Selected excerpts from participant comments on the provided tips (disrupted configuration).

Many participants felt that the human tip was more accurate since it better matched their intuition, and disagreed with tips that they found counter-intuitive. Some participants even found the human tip to be counter-intuitive since it did not match their intuition from the fully-staffed scenario in prior rounds (i.e., it asks the server to cook once instead of not at all); this result is matched by a post-game survey question for the normal configuration where participants were asked to imagine how their strategy would change in the under-staffed scenario (see Appendix B.5). As a consequence, compliance and performance suffered. Importantly, we observe that even participants who successfully understood our algorithm's tip (and viewed it favorably at the end) claimed that they did not comply with the tip in earlier rounds of the understaffed scenario. Rather, they needed time to experiment with and without the tip in order to learn its value.

Our results suggest that, to achieve high compliance, it is not sufficient for the participant to just understand the action suggested by the tip (a major focus of the literature on interpretable machine learning); they also have to believe that the suggested action will help improve performance, and be given sufficient time to learn how to correctly incorporate the tip into their workflow.

For robustness, we also performed these sentiment analyses using two natural language processing approaches, VADER and BERTweet, finding qualitatively similar insights. First, we use a lexicon and rule-based sentiment analysis tool called VADER (Hutto and Gilbert 2014). VADER provides an API *polarity_scores* that returns four values for each set of texts, including *pos* for positive sentiment, *neg* for negative sentiment, *neu* for neutral sentiment, and a compound score. The compound score is normalized to be between -1 (most extremely negative) and $+1$ (most extremely positive). According to the developer of VADER, among these four scores, the compound score is "the one most commonly used for sentiment analysis by most researchers, including the authors."

However, in practice, although VADER generates a reasonable score for most inputs, it does not perform as well for more complex content. For example, it might classify a response as positive because it sees the word "like" repeatedly, when the word is actually being used as a preposition or conjunction. To address this issue, we instead consider a pre-trained BERT model to classify sentiments. Specifically, we used BERTweet Base Sentiment Analysis, a RoBERTa model trained on English tweets (Pérez et al. 2021). We used the pipeline API provided by Hugging Face's Transformers library. Because the model returns three separate scores: *Positive*, *Neutral*, and *Negative*, to get a final compound score as we had before, we calculated $Positive - Negative$.

Table C.8 exhibits all the scores from our VADER and BERTweet analyses using the responses among participants in the disrupted configuration of Phase II. We find that our results are highly consistent with our initial approach using the human coder—i.e., the human tip consistently has higher positive, lower negative, and higher compound scores than our algorithm's tip across both approaches, suggesting that participants consistently perceive our algorithm's tip to be less favorable.

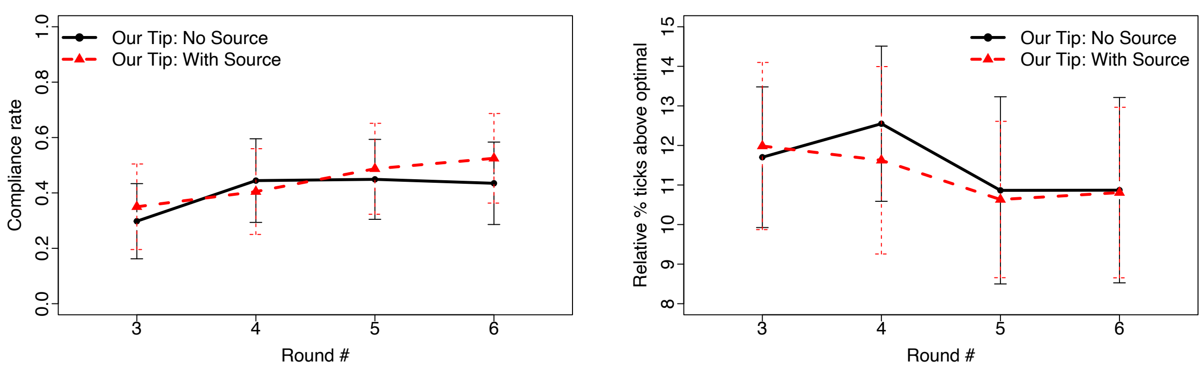
Disrupted	VADER+	VADER	VADER−	VADER	B+	B	B−	BERTweet
Algorithm	0.0922	0.8515	0.0563	0.0256	0.2113	0.4712	0.3176	−0.1063
Human	0.1042	0.8601	0.0358	0.1305	0.2473	0.6100	0.1427	0.1046
Baseline	0.0979	0.8339	0.0682	0.0600	0.1315	0.5029	0.3657	−0.2342

Table C.8 Positive, neutral, negative, and compound sentiment scores from VADER and BERTweet, using Phase II's disrupted configuration results. $+$, $,-$ **refer to positive, neutral, and negative scores, respectively.**

C.6. Trust in Algorithms and Compliance

While our experiments thus far did not reveal the *source* of the tip (i.e., whether it is generated by an algorithm), workers may be able to infer this information in real-world contexts, potentially resulting in algorithm aversion (Dietvorst et al. 2018)—i.e., where humans are mistrustful of algorithmic advice. To this end, we perform a pilot study in the disrupted configuration to evaluate the impact of algorithm aversion on compliance. We randomly assign participants into one of two conditions: "Our Algorithm: No Source" and "Our Algorithm: With Source". The "Our Algorithm: No Source" condition is identical to the "Our Algorithm" condition in our main study, i.e., the participant is shown "Tip: Server should cook twice" during the under-staffed rounds. In contrast, in the "Algorithm: With Source" condition, the participant is instead shown: "Tip from AI Algorithm: Server should cook twice. The AI algorithm analyzes past players' strategies and chooses the best tip that would help improve your performance."

We recruited 200 participants via AMT, of which 90 successfully completed the study and passed all comprehension and attention checks. We find that there are no statistically or economically significant differences in the results between the two conditions according to compliance rate (Figure C.5a) or final round performance (Figure C.5b). Providing the source of the tip in fact has a directionally *positive* impact on compliance, suggesting that it is unlikely that we would observe algorithm aversion. One potential explanation is that, given the complex nature of the task, knowing that the tip came from an algorithm could increase humans' likelihood in adopting the tip similar to the phenomenon of algorithm appreciation documented by Logg et al. (2019). Note that the final round performance is essentially identical whether or not the source was provided.



(a) Compliance Rate with Our Algorithm's Tip (b) Performance over Time

Figure C.5 Source of Tip. Participant compliance with our algorithm's tip ("Server should cook twice") (left) and participant performance (right) whether the information about the source of the tip was provided.

C.7. Details on Compliance Intervention Experiment

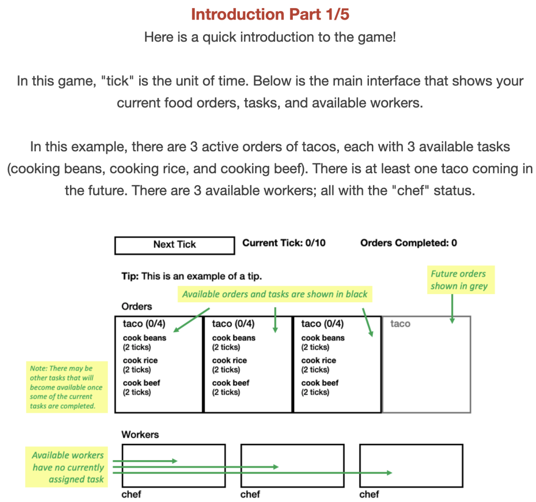
To disentangle compliance from performance, we consider several interventions that have been shown to improve compliance in prior work. First, we consider financial incentives, which have been demonstrated to improve human adoption of machine-generated advice (e.g., Giuffrida and Torgerson 1997, Homonoff 2018, Beshears et al. 2021, Milkman et al. 2021). Another popular intervention in practice is to use social information; past research has shown that information about social norms can change human behavior and improve compliance with guidelines (Benjamin et al. 2010). For example, residents consumed less energy after learning that their neighbors had better energy consumption ratings (Allcott 2011). Finally, one of the challenges we identify is that participants have difficulty understanding our tip. Thus, inspired by the strategy of curriculum learning in machine learning (Bengio et al. 2009), we consider a strategy that first provides a simpler version of our tip. Focusing on the disrupted configuration, we investigate the following four interventions in the under-staffed rounds of the disrupted configuration:

1. "Pay" condition: For rounds 3 and 4 (first two under-staffed rounds), we pay the participant the maximum pay for each round *if* they successfully complied with the tip (i.e., server cooked twice). The pay scheme returns to the original performance-based one in rounds 5 and 6.
2. "Social" condition: We add the following to the tip for all four under-staffed rounds—"While this tip may appear counter-intuitive, the majority of best players adopted this rule, enabling them to achieve the optimal performance of 34 ticks."
3. "Pay-Social" condition: Participants receive both the Pay and Social interventions.
4. "Curriculum" condition: In round 3 (e.g., the first disrupted round), we present the Human tip from the original study ("Server should cook once") instead of our algorithm's tip. Then, in rounds 4 through 6, we present our algorithm's tip ("Server should cook twice"). The rationale of this intervention is to slowly move the participant's strategy from not letting the server cook any burgers to something in between (letting the server cook once) before telling them the more counter-intuitive algorithm tip (letting the server cook twice).

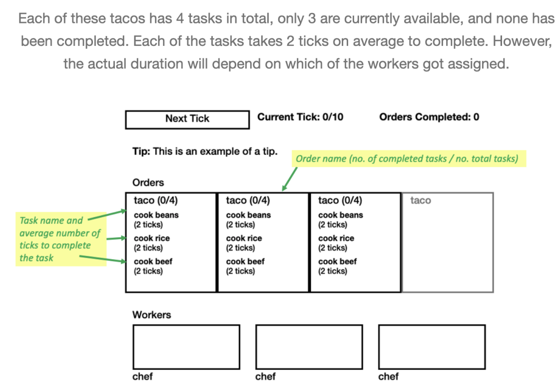
We recruited 1,967 participants via AMT, of which 1,496 successfully completed the study and passed all the comprehension and attention checks. Participants were randomly assigned into one of five conditions: "Tip Only" (identical to the Algorithm arm in the original study), "Pay", "Social", "Pay-Social", and "Curriculum" interventions.

Appendix D: Screenshots of Our Virtual Kitchen Management Game

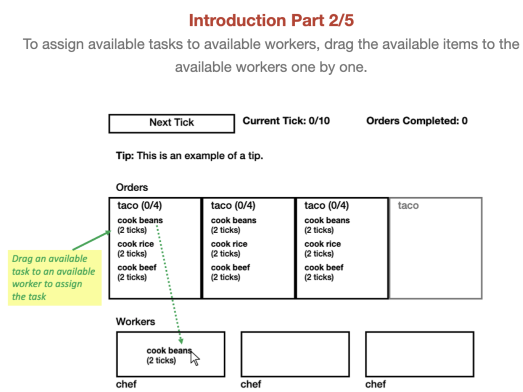
Finally, we provide screenshots to illustrate our experimental design. Figures D.1 and D.2 show the introduction to the task shown to participants explaining various concepts in the game. Figures D.3 and D.4 show instructions for the fully-staffed and understaffed scenarios, respectively, shown to participants. Finally, Figure D.5 shows the payment information shown to the participants.



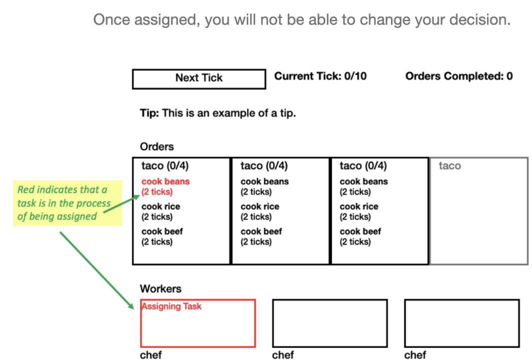
(a) Introduction to the interface



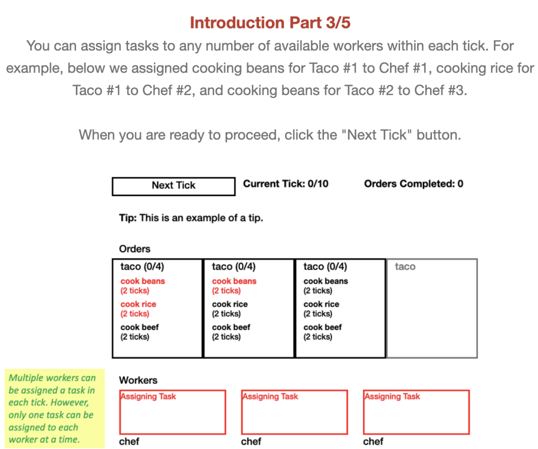
(b) Introduction to the subtasks



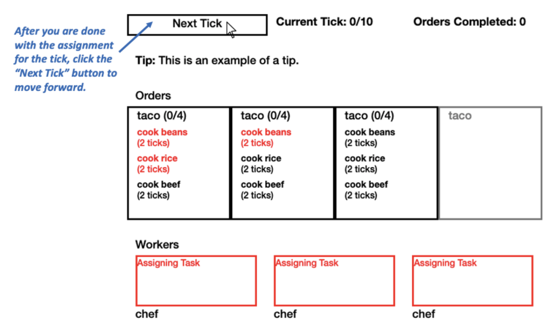
(c) Introduction to task assignment



(d) Introduction to task assignment (cont.)

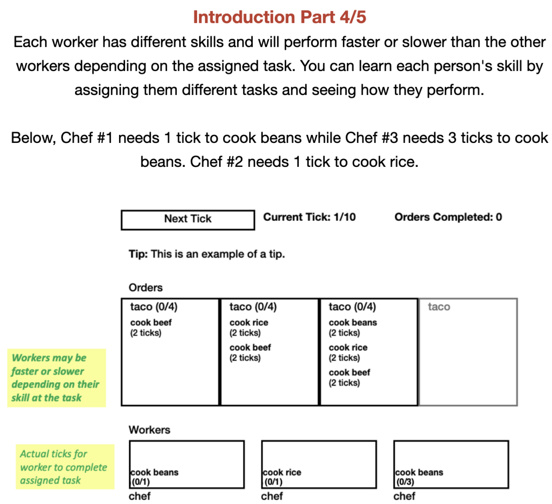


(e) Introduction to task assignment (cont.)

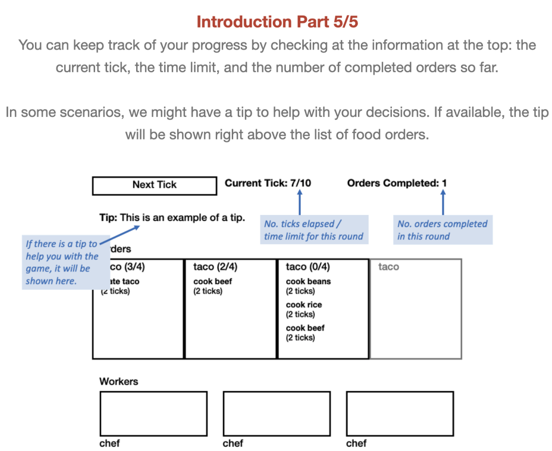


(f) Introduction to task assignment (cont.)

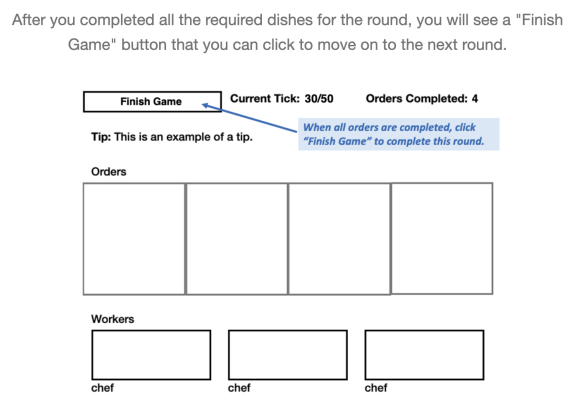
Figure D.1 **Screenshots of the game introduction.**



(a) Introduction to workers' skill levels



(b) Introduction to the tip



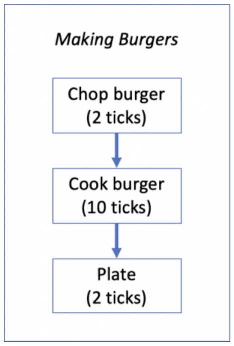
(c) Introduction to round completion

Figure D.2 Screenshots of the game introduction (continued).

Round 1: Burger Queen

In this round, we will be serving a new dish: **burgers**.

- Making a burger involves chopping meat, cooking the burger, and plating the final dish (see diagram below).
- You will have access to 3 different workers for the next few rounds: Chef, Sous-Chef, and Server. Remember, each worker is faster at different tasks.



Goal

- You have 50 ticks to serve 4 burgers as fast as possible for the highest pay.
- Most players finish in 26 ticks in this round, but our best players finish in 20 ticks.
- **You must serve all orders within the time limit to be qualified for payment.**

Bonus

- You will receive an additional \$0.15 for finishing within 26 ticks,
- an additional \$0.20 for finishing within 22 ticks,
- and an additional \$0.40 for finishing within 20 ticks.

Remember:

- The key to serving burgers fast is a well-managed kitchen! Make sure to play to your workers' strengths.
- When you're done assigning tasks for the tick, **press "Next Tick" button to move forward.**

(a) Burger's subtasks and available workers (b) Goal, incentives, and reminder

Figure D.3 Screenshots of the instructions for the fully-staffed scenario.

Round 3: Burger Queen

Unfortunately, the Chef is on vacation during this round. **(Who travels these days...) Now you only have 2 workers in the kitchen.**

Goal

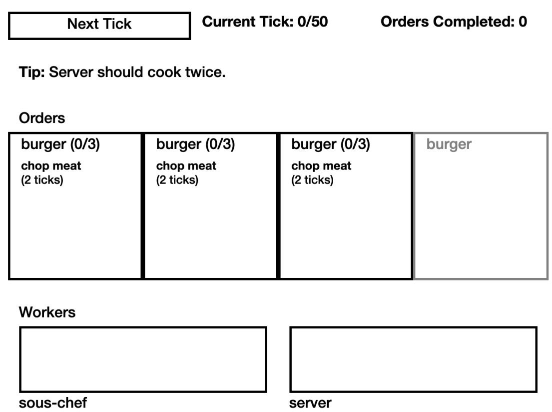
- You have 50 ticks to serve 4 burgers as fast as possible for the highest pay.
- Most players finish in 38 ticks in this round, but our best players finish in 34 ticks.
- **You must serve all orders within the time limit to be qualified for payment.**

Bonus

- You will receive an additional \$0.15 for finishing within 38 ticks,
- an additional \$0.20 for finishing within 36 ticks,
- and an additional \$0.40 for finishing within 34 ticks.

Remember:

- The key to serving burgers fast is a well-managed kitchen! Make sure to play to your workers' strengths.
- When you're done assigning tasks for the tick, **press "Next Tick" button to move forward.**



(a) Updated instructions following the in-game disruption (b) Game interface (with the algorithm tip)

Figure D.4 Screenshots of the instructions for the understaffed scenario.

Round 1: 24 ticks ---> **\$0.15**

Nice job! Think you can do better? Here's a chance to try again!

Summary of Bonuses

Round 1: 20 ticks ---> **\$0.75**
Round 2: 20 ticks ---> **\$0.75**
Round 3: 36 ticks ---> **\$0.35**
Round 4: 38 ticks ---> **\$0.15**
Round 5: 39 ticks ---> **\$0**
Round 6: 34 ticks ---> **\$0.75**

Base Pay + Bonuses: \$3.35.
Nice job!

(a) Individual round pay information (b) Summary of total pay

Figure D.5 Screenshots of the pay information.

References for Online Appendix

Allcott H (2011) Social norms and energy conservation. *Journal of Public Economics* 95(9-10):1082–1095.

Bengio Y, Louradour J, Collobert R, Weston J (2009) Curriculum learning. *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Benjamin DJ, Choi JJ, Strickland AJ (2010) Social identity and preferences. *American Economic Review* 100(4):1913–28.

Beshears J, Lee HN, Milkman KL, Mislavsky R, Wisdom J (2021) Creating exercise habits using incentives: The trade-off between flexibility and routinization. *Management Science* 67(7):4139–4171.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64(3):1155–1170.

Giuffrida A, Torgerson DJ (1997) Should we pay the patient? review of financial incentives to enhance patient compliance. *BMJ* 315(7110):703–707.

Homonoff TA (2018) Can small incentives have large effects? the impact of taxes versus bonuses on disposable bag use. *American Economic Journal: Economic Policy* 10(4):177–210.

Hutto C, Gilbert E (2014) Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the international AAAI conference on web and social media*, volume 8, 216–225.

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151:90–103.

Milkman KL, Gromet D, Ho H, Kay JS, Lee TW, Pandiloski P, Park Y, Rai A, Bazerman M, Beshears J, et al. (2021) Megastudies improve the impact of applied behavioural science. *Nature* 600(7889):478–483.

Pérez JM, Giudici JC, Luque F (2021) pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks. *arXiv preprint arXiv:2106.09462* .