Cognition 264 (2025) 106200

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

People evaluate idle collaborators based on their impact on task efficiency

Elizabeth Mieczkowski [a,*], Cameron Turner [a,b], Natalia Vélez [b], Thomas L. Griffiths [a,b]

[a] Department of Computer Science, Princeton University, 35 Olden St, Princeton, 08540, NJ, USA
[b] Department of Psychology, Princeton University, Peretsman Scully Hall, Princeton, 08540, NJ, USA

ARTICLE INFO

Dataset link: https://github.com/emieczkowski/IdleCollaborators

Keywords:
Collaboration
Distributed systems
Collective intelligence
Amdahl's Law
Social loafing
Social norms

ABSTRACT

Humans collaborate to improve productivity, but when is it acceptable for a collaborator to remain idle? Theories from distributed computer systems suggest that, depending on the task structure, division of labor leads to diminishing returns in efficiency as group size increases. We examine whether people are aware of these limitations to collaboration, and how considerations of task efficiency may affect the perceived acceptability of idleness, the withholding of effort during collaborative tasks. Across four experiments ($N = 1,124$), participants saw scenarios where a single collaborator remained idle while other group members washed dishes, prepared a salad, or created flashcards. We manipulated task structure by varying the number of guests (group size), the amount of work to be done (workload), and the number of tools available to do it (environmental bottlenecks), which each constrain how much faster the group could have finished the task if the idle agent had contributed. Participants judged idleness as more acceptable when the idle agent's contributions would have a smaller effect on task efficiency. These judgments were best captured by a variant of Amdahl's Law, a theory from distributed systems that predicts the idle agent's potential impact by integrating group size, workload, and bottlenecks, compared to simpler heuristic models that consider a subset of these factors. Together, our findings lay the groundwork to study human collaborations as natural distributed systems.

1. Introduction

Humans work together to accomplish goals beyond the capabilities of a single person. By coordinating actions as a group, we are able to leverage collective effort and overcome individual constraints on time and resources (Griffiths, 2020; Vélez, Christian, Hardy, Thompson, & Griffiths, 2023). Groups of people are thus motivated to collaborate based on the underlying belief that "many hands make light work"; as we pool shared resources, skills, and effort, we should benefit from an increase in collective productivity (Mao, Mason, Suri, & Watts, 2016). However, figuring out how labor should be allocated across collaborators is not trivial. How many collaborators do we actually need in order to complete tasks efficiently?

We propose that theories from distributed computer systems provide a theoretical framework for understanding the challenges faced by groups of people as they transition from working individually to collaborating (Vélez et al., 2023). Splitting tasks among many processors, as in a computing cluster, can enable us to solve computational problems that outstrip the capabilities of any one machine, such as capturing images of black holes or training complex AI models. However, this process introduces challenges similar to those encountered by human collaborators, such as coordinating concurrent operations, agreeing on shared states, and handling failures like machine crashes or communication delays (Almasi & Gottlieb, 1994; Baer, 1973; Kshemkalyani & Singhal, 2011). For decades, research in distributed systems has formalized these coordination problems and their solutions, providing insights through theories like Amdahl's Law, which predicts an upper bound on performance gains from increasing group size. Depending on the number of processors already in use and the structure of the task, adding more processors may have little impact on performance (Amdahl, 1967; Cassidy & Andreou, 2011; Gustafson, 1988; Hill & Marty, 2008). Applied to human collaborations, Amdahl's Law offers a principled way to analyze how much faster a task could be completed by distributing it across more collaborators.

Distributed systems motivate the idea that, depending on the task, there may be diminishing returns to division of labor. For instance, consider a dinner party where all guests have finished eating and it is time to clean the dishes. Should an additional person try to contribute, or is it acceptable for them to remain idle? If task efficiency influences how people evaluate collaborations, then the variables predicting speedup in

* Corresponding author.
E-mail addresses: emiecz@princeton.edu (E. Mieczkowski), c.rouse.turner@princeton.edu (C. Turner), nvelez@princeton.edu (N. Vélez), tomg@princeton.edu (T.L. Griffiths).

https://doi.org/10.1016/j.cognition.2025.106200
Received 12 August 2024; Received in revised form 28 May 2025; Accepted 1 June 2025
Available online 11 July 2025

distributed systems should affect how people perceive idleness. Drawing on principles from Amdahl's Law (Amdahl, 1967), the degree to which a task can be parallelized between collaborators should depend on: (1) the number of people (group size), (2) the amount of work to be completed (workload), and (3) serial limitations that prevent additional members from effectively contributing (bottlenecks). Bottlenecks can be environmental, such as spatial constraints or physical resources, or cognitive, such as communication delays or interdependencies between task components that require synchronization (Almaatouq, Alsobay, Yin, & Watts, 2021). If there is a low workload, such as a limited number of plates to clean, few guests can help before there is nothing left to do. Similarly, if there are many dirty plates but only one sink, an extra helper will not reduce washing time due to the physical bottleneck. Examples of these definitions are illustrated in Fig. 1. In such cases, idleness may not be a failure to help, but instead a rational collaborative strategy. Individuals might refrain from exerting effort when they perceive that their involvement will not significantly enhance overall task efficiency.

Many psychological theories suggest that people evaluate events not only based on what actually happened, but by considering what could have happened instead. Counterfactual Simulation Models provide valuable insights into how individuals mentally simulate alternative outcomes to infer causality and evaluate decisions (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015; Gerstenberg, Lagnado, et al., 2023; Xiang, Landy, Cushman, Vélez and Gershman, 2023). A key challenge in applying these models is determining which counterfactuals are most relevant for judgment. In the context of idleness evaluations, a particularly important counterfactual concerns task completion time: how much faster would the work have been completed if the idle agent had contributed? From this perspective, judgments of idleness may reflect an implicit counterfactual comparison. If the task would have been completed faster with the idle agent, they should be judged more negatively (Gerstenberg et al., 2018; Wu & Gerstenberg, 2024; Wu, Sridhar, & Gerstenberg, 2023). This approach complements existing work on counterfactual reasoning by highlighting how structural factors in the environment influence potential contributions. While counterfactual models effectively capture how people evaluate alternative scenarios when judging contributions, distributed systems theory helps explain why specific tasks and environmental factors make certain counterfactuals more relevant than others. This framework provides concrete mechanisms for understanding how resource constraints and coordination requirements shape the potential impact of individuals' effort in groups (Gerstenberg & Lagnado, 2010).

However, task efficiency may not be the only explanation for how idleness should be judged during collaborations. Our model based on distributed systems theory offers a comprehensive approach to understanding how group size, workload, and environmental bottlenecks jointly influence the perceived acceptability of idleness. Instead of evaluating task efficiency using multiple variables, people might rely on simpler heuristics that consider only a subset of these factors. One possibility is that people evaluate idle collaborators based on the current group size. Different psychological theories offer competing predictions about whether group size increases or decreases the perceived acceptability of idleness. On one hand, idle collaborators may be judged more harshly in larger groups if their evaluations are influenced by social norms (Cialdini, Kallgren, & Reno, 1991; Hackman & Morris, 1975; Roberts, Gelman, & Ho, 2017; Schmidt & Tomasello, 2012; Turner, Nielsen, & Collier-Baker, 2014). As more people contribute to the clean-up, there will be heightened normative pressure on those remaining seated to participate in order to avoid being perceived as uncooperative in the eyes of the productive group members (Mas & Moretti, 2009). On the other hand, idle collaborators may be judged more leniently in larger groups if evaluations are based on expectations of social loafing, or the tendency for people to expend less effort when working together (Ingham, Levinger, Graves, & Peckham, 1974; Petty, Harkins, Williams, & Latane, 1977; Ringelmann, 1913; Steiner, 1972). People may judge idleness as more acceptable in larger groups because everyone has a tendency to 'loaf' and do the same. Another possibility, motivated by work in management science, is that individuals will tailor the amount of effort they exert depending on workload (Bruggen, 2015). If workload is perceived as too high or low, individuals might be less motivated to contribute, making idleness more acceptable.

Here, we tested whether people's evaluations of idle collaborators are sensitive to task efficiency, as predicted by a modeling framework inspired by distributed computer systems. We examine how people judge the social acceptability of idleness in an experimental paradigm in which all group members except for one contribute to a collaborative task. In Experiment 1, we hypothesize that people consider both workload and group size when evaluating idleness. If participants base collaborative judgments on task efficiency, then they should consider idleness to be more acceptable in cases where an additional collaborator will have little impact, such as when workloads are low and group sizes are large. In Experiment 2, we predict that bottlenecks also impact idleness judgments. Participants should judge idleness to be less acceptable when there are fewer environmental constraints to working collaboratively, such as having access to a larger number of available workstations and tools. To formalize these predictions, we develop a mathematical model of task efficiency in collaborative groups inspired by Amdahl's Law from distributed computer systems (Amdahl, 1967). Consistent with our model, Experiments 1 and 2 reveal qualitative similarities between participants' judgments and the predicted speedup gained if the idle agent had contributed. In Experiment 3, we directly test this relationship, finding that participants' estimates of potential task speedup strongly correlate with evaluations of idleness. Experiment 4 replicates these findings in a novel scenario that removes spatial affordances and contextual norms associated with dinner parties, reinforcing the robustness of our results. Together, these results suggest that people evaluate idleness based on how much a group member *could have* contributed, consistent with basic principles of task efficiency inspired by distributed systems.

2. Modeling the impact of idleness with distributed systems efficiency

Theories from distributed systems offer a framework that can be adapted to model the acceptability of idleness based on task efficiency. Parallelization enables processors to carry out complex computations simultaneously, leading to immense improvements in performance. However, coordinating multiple processors to tackle the same operations in parallel produces tremendous challenges (Almasi & Gottlieb, 1994; Baer, 1973; Kshemkalyani & Singhal, 2011). Distributed systems theory highlights the potential disadvantages of allocating tasks across multiple processors and provides general principles for improving efficiency.

We develop a mathematical model to predict the acceptability of idleness during collaborative tasks depending on task efficiency. If people are influenced by task efficiency when evaluating idleness, then they should find it more acceptable to withhold effort when their contribution has little impact on speeding up task completion. To explain our derivation, we return to our real-world example of a dinner party. After dinner, several guests begin cleaning the dishes, which involves two subtasks: (1) Guests must first carry all of the plates to the counter, and (2) then wash them. How much of an impact can an additional guest have on speeding up clean-up? This difference should partially depend on the number of plates that need to be cleaned, or the *workload*. An extra guest can have a greater marginal impact if there are twenty plates to be cleaned compared to only two. Furthermore, this guest can contribute more if *group size* is currently smaller, such as if only two people are working on the task compared to ten. Finally, this guest's potential impact is limited by the parallelizability of each subtask due to *bottlenecks*. Environmental factors, such as spatial constraints or a limited number of workstations and tools, create these

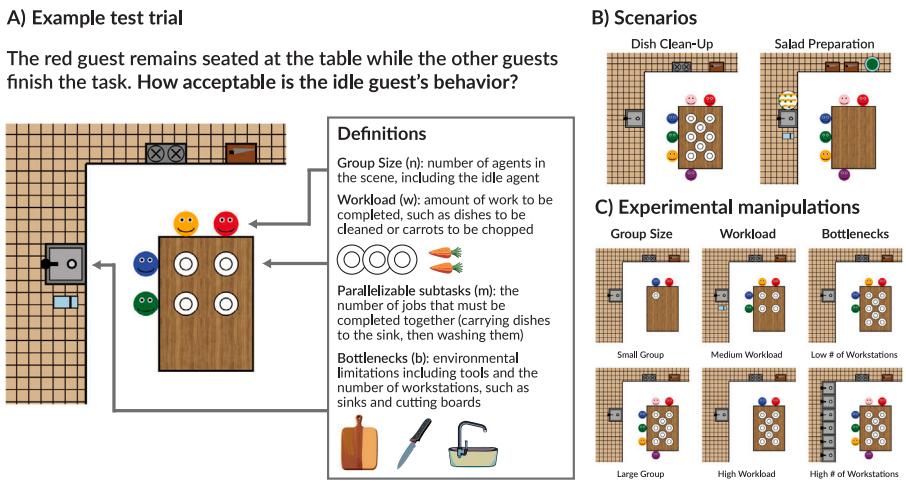

Fig. 1. Definitions and example stimuli. (A) Example trial: Participants were told that all guests except for the red agent actively participate in the task, and they were asked to evaluate how acceptable it is for the red agent to remain idle. (B) Scenario types: Agents worked together to wash dishes (Dish Clean-Up) or make a salad (Salad Preparation). (C) Experimental manipulations: We manipulated group size, workload, and environmental bottlenecks like the number of workstations, which each affect the red agent's impact on task efficiency. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

bottlenecks and prevent successful division of labor. In this case, the primary bottleneck is the number of sinks. If there is only one sink, and there are already several guests crowded around it, then another guest cannot speed up this subtask. In summary, our model depends on a combination of (1) **workload** w, (2) **group size** n, (3) the number of **parallelizable subtasks** m, and (4) the presence of environmental **bottlenecks** b_i that limit the number of agents who can successfully share work within a subtask i. Fig. 1A defines each of these variables and shows how they were instantiated in our stimuli.

2.1. Derivation from original Amdahl's Law

Amdahl's Law is a formula that predicts the speedup expected when a computation is executed across more processors, based on the proportion of subtasks that can be parallelized (Amdahl, 1967; Cassidy & Andreou, 2011; Hennessy & Patterson, 2011; Hill & Marty, 2008). The original version of Amdahl's Law assumes that subtasks can fall into two categories. The first proportion of subtasks f can be fully parallelized (e.g., carrying dishes to the counter). The second proportion of subtasks $1 - f$ must be performed serially by a single processor (e.g., washing dishes one at a time). Given a fixed workload, Amdahl's Law predicts the theoretical speedup S expected by distributing a task among multiple processors:

$$S(s) = \frac{1}{(1-f) + \frac{f}{s}} \tag{1}$$

where s is a speedup factor, representing the time saved in the parallelizable proportions of the task. The key insight from this formula is that, as group size increases, improvements in task efficiency are constrained by the workload and the proportion of the task that must be performed serially. Amdahl's Law points to an inherent limit to how much increasing group size can enhance task efficiency. It is only beneficial to add a processor when there is unfinished work and additional processors will reduce the time to completion.

We extend this model in two ways. First, we consider more than two proportions of a task, each with its own degree of parallelizability. In our real-world example, subtasks are not strictly collaborative or independent; some subtasks may benefit from varying numbers of collaborators. Any number of guests may be able to bring their own plates to the sink, whereas two or three guests may dry them depending on the number of towels. Second, we consider factors that impact speedup s in real-world settings. In distributed computing, s is often calculated based on the number of discrete operations that can be run simultaneously, such as counting the elements of a certain type in a list or performing matrix multiplications (Hennessy & Patterson, 2011). However, for human collaborations in real-world tasks, defining the number of actions and amount of effort required for each subtask is less straightforward.

Applying Amdahl's Law to human groups motivates two important hypotheses for understanding collaboration. First, adding group members to a task may not improve efficiency, depending on the workload and environmental bottlenecks that limit parallelizability. Second, if people are aware of these limitations, idleness may be a rational response to the diminishing returns expected from increasing group size.

2.2. Task efficiency model: Evaluating the impact of idle collaborators

To study the impact of idleness in human collaborations, we developed a *task efficiency model* inspired by Amdahl's Law (Amdahl, 1967; Cassidy & Andreou, 2011; Hennessy & Patterson, 2011; Hill & Marty, 2008). Our model builds on the original form of Amdahl's Law (Eq. (1)) by predicting the impact of adding a single collaborator on task completion times and using this prediction to assess the acceptability of idleness. The predicted speedup that could be achieved if one more agent contributed to the task is given by:

$$S(n,b) = \frac{1}{(1 - \sum_{i=1}^{m} f_i) + \sum_{i=1}^{m} \frac{f_i}{s(n,b_i)}} \tag{2}$$

Here, m is the number of parallelizable subtasks. We assume that each subtask contributes equally to the overall task, with f_i representing the fraction of the task comprised by a parallelizable subtask i. Therefore, $\sum_{i=1}^{m} f_i = 1$ if all subtasks are parallelizable. Conversely, $1 - \sum_{i=1}^{m} f_i$ represents the serial proportion of the task. For example, in a dish clean-up scenario with only one sink, there are two distinct subtasks: clearing the plates from the table, and washing them. Only the first subtask is parallelizable, so $m = 1$. We approximate this subtask as compromising $f_1 = 0.5$ of the overall completion time. Therefore, the second subtask, which must be performed serially because there is only one sink, comprises $1 - f_1 = 0.5$ of the task. Because we currently consider static descriptions of real-world tasks, we do not have an exact quantification of the total number of actions needed for each subtask. Thus, we approximate the proportion of time allocated to each subtask by splitting the total task completion time evenly across subtasks. In the Discussion, we will consider ways to divide subtasks more precisely in dynamic, first-person tasks.

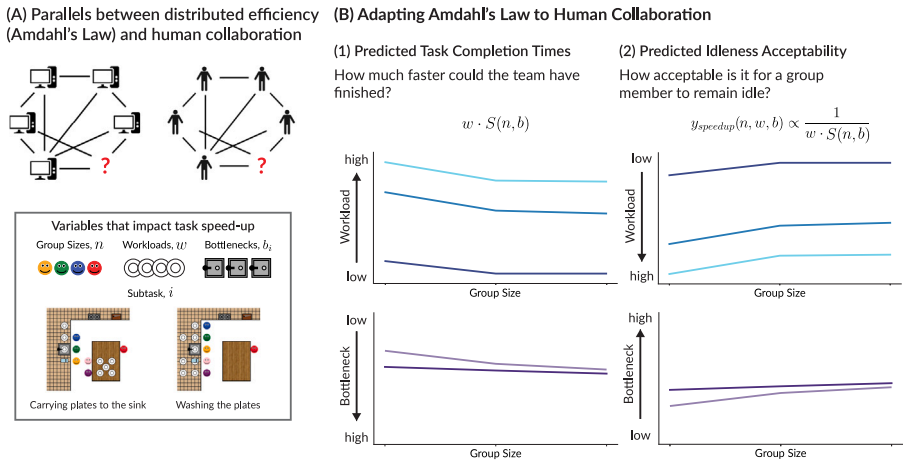

Fig. 2. Schematic of task efficiency model, based on Amdahl's Law from distributed systems. (A) Parallels between task efficiency in distributed computer systems (Amdahl's Law) and human collaboration. We consider including an additional team member in a collaborative task to be similar to scaling system size in distributed systems (top). We tailor our model to account for variables that impact human collaborative efficiency, namely group sizes, workloads, and bottlenecks (bottom). In our model, serial proportions of a task can only be performed by one collaborator (e.g., washing the dishes one at a time) due to environmental bottlenecks. (B) Task efficiency model predictions: Predicted task completion times and idleness acceptability judgments across different group sizes, workloads, and bottlenecks (i.e., number of workstations). We predict that idleness acceptability is inversely proportional to task speedup. Predicted speedup is presented on a log scale.

For each parallelizable subtask i, $s(n, b_i)$ estimates the speedup from adding an additional nth agent, given there are already $n-1$ agents working on a task with b_i workstations:

$$s(n, b_i) = \frac{\min(n, b_i)}{\min(n-1, b_i)} \tag{3}$$

To illustrate, consider a scenario with five agents and two sinks ($n = 5$, $b_2 = 2$). Adding another agent has no impact due to the limited number of workstations. Conversely, with five agents and six sinks ($n = 5$, $b_2 = 6$), then recruiting another agent improves task completion due to the availability of workstations.

Amdahl's Law predicts the speedup obtained from adding processors, given a fixed workload. To facilitate comparisons across different workload levels, we introduce a workload parameter w. We then compute overall task efficiency by scaling the speedup $S(n, b)$ by w. We consider idleness acceptability to be inversely proportional to speedup, representing the "slow-down" that occurs when the idle agent does not contribute. Thus, we predict idleness acceptability given speedup efficiency, $y_{\text{speedup}}(n, w, b)$ (depicted in Fig. 2), to be

$$y_{\text{speedup}}(n, w, b) = \frac{1}{w \cdot S(n, b)} \tag{4}$$

3. Alternative hypotheses

Our key hypothesis is that idleness is more acceptable when the idle agent's contributions would have had less impact on task completion time. The task efficiency model computes this impact by incorporating information about group size, workload, and bottlenecks. However, there are several alternative possibilities. Instead of combining all three task features, people might rely on simpler heuristics that consider only a subset. In particular, people's judgments may be primarily based on group size; it may be more acceptable because there are more hands to share the load (*social loafing model*). Alternatively, group size may make it less acceptable to remain idle in larger groups due to stronger normative pressure to contribute (*normative pressure model*). Alternatively, people may base their judgments solely on workload, with idleness seen as less acceptable when there is more work to do (*workload model*). We describe each of these alternative hypotheses below.

3.1. Social loafing models

Individuals frequently exert less effort when working in larger groups, a phenomenon known as social loafing. This phenomenon, first demonstrated by Ringelmann (1913), showed that people each contributed less effort to a rope-pulling task when pulling together than when pulling individually (Kravitz & Martin, 1986; Ringelmann, 1913). Many accounts explain social loafing as a failure to coordinate effort within a group, and this phenomenon has been replicated across a wide range of populations and tasks, both physical and cognitive (Ingham et al., 1974; Karau & Williams, 1993; Petty et al., 1977; Simms & Nichols, 2014). This research indicates that the severity of social loafing increases with group size (Littlepage, 1991; Ringelmann, 1913; Steiner, 1972). Social loafing may worsen as group size increases due to compounding coordination challenges, greater dispensability and unidentifiability of individual contributions, and increased expectations of peer loafing (Jackson & Harkins, 1985; Kerr & Bruun, 1983; Williams, Harkins, & Latané, 1981).

Although these theories differ in their explanations of *why* group members decrease individual outputs, they generally agree that effort depends on perceived group size. As the group grows larger, each member contributes less effort, and the discrepancy between the potential and actual effort exerted by the group becomes more pronounced (Olson, 1971).

If people use similar considerations to evaluate others' idleness as they do to calibrate their own effort, they may judge idleness as more acceptable in larger groups since there are more collaborators available to carry the load. Therefore, we can summarize these theories into a heuristic which predicts that idleness judgments given n group members will be:

$$y_{\text{social loafing}}(n) = l(n) \tag{5}$$

where $l(n)$ is a function of group size that can take various forms depending on the underlying psychological theory. We use both a linear and a nonlinear model to capture this relationship, given that social loafing has been shown to follow either pattern in terms of its effect on performance (Littlepage, 1991; Ringelmann, 1913; Steiner, 1972). Specifically, we consider a linear model $l(n) = n$, and a nonlinear model $l(n) = \log(n)$.

3.2. Normative pressure models

Alternatively, group size may have the opposite effect on people's evaluations of idleness. Idleness may be less acceptable in larger groups due to greater normative pressure on each individual to conform to the majority. People are sensitive to conformity; they both tune their own behaviors to what the rest of the group is doing (Asch, 1940; Flynn, Turner, & Giraldeau, 2018; Gerard, Wilhelmy, & Conolley, 1968; Turner et al., 2014) and evaluate other people's behaviors based on how well they align with the group (Bear & Knobe, 2017; Cialdini & Goldstein, 2004; Cialdini et al., 1991). Thus, as group size increases and more guests help with the dishes, people may evaluate the idle collaborator's actions more negatively simply because it is increasingly at odds with the actions of the rest of the group.

If people judge idleness based on descriptive norms about how many people are actively engaged, then it may be less acceptable in larger groups, where there is greater normative pressure to act. In this case, idleness judgments given n group members will be predicted by:

$$y_{\text{norms}}(n) = \frac{1}{n} \tag{6}$$

3.3. Workload models

Both the social loafing and normative pressure models base their predictions solely on group size, but people might be primarily concerned with workload. Research in management science suggests that workload heavily impacts the amount of effort individuals exert during a task. Greater workload can have a positive impact on effort, leading to increased motivation and performance (Siswanto, Supriyanto, Ni'mah, Asnawi, & Wekke, 2019). Other work finds an inverted U-shaped relationship between workload and individual performance, where employee outputs increase up to a certain workload and then decrease (Bruggen, 2015).

If people base their collaborative judgments solely on workload, then they may judge idleness as less acceptable when there is more work to be done. If this is the case, idleness judgments based on workload w can be roughly described as:

$$y_{\text{workload}}(w) = \frac{1}{w} \tag{7}$$

Alternatively, a more sophisticated model might capture the empirical, non-monotonic relationship between workload and performance, such as:

$$y_{\text{effort}}(w) = -w^2 \tag{8}$$

3.4. Model summary

Each of the alternative hypotheses described above evaluates idleness based on a single task dimension: the social loafing and normative pressure models evaluate idleness solely using group size n, while the workload model evaluates idleness using workload w. In contrast, our task efficiency model uniquely combines considerations of group size, workload, and environmental bottlenecks b within a single theoretical framework.

4. Experiment 1: Effects of workload and group size

We present a paradigm for evaluating whether people take into account task efficiency when judging whether an additional group member should contribute to a collaborative task. In particular, if participants are sensitive to task efficiency, then they should consider both workload and group size when evaluating the acceptability of idleness.

4.1. Methods

Participants were presented with a series of dinner party scenarios. In each scene, all agents except for one idle guest were actively engaged in a task. Participants were asked to evaluate the acceptability of this agent's idleness.

4.2. Preregistration

All methods, including the experimental design, data collection procedures (e.g., sample size, participant recruitment, inclusion/exclusion criteria), and our primary subject-level analysis (a linear mixed-effects model comparing observed data to our hypothesized task efficiency model), were preregistered prior to data collection at https://aspredicted.org/8bj9q.pdf. These preregistered elements were intended to test our key hypotheses regarding the effect of task efficiency on judgments of idle behavior. We did not preregister the comparison to additional baseline models, as those comparisons are ablations of the full task efficiency model that help us interpret our theoretical results. Thus, comparisons with alternative models are exploratory. The experiment was performed with IRB approval (IRB # 15959); all participants provided informed consent prior to the experiment.

4.3. Participants

We recruited 300 English-speaking adults from the Prolific platform in exchange for compensation ($2.40 for a 10–12 min experiment). Fourteen participants who failed an attention check were excluded from analysis, resulting in a total of 286 participants (87 male, 196 female; mean age = 38.8, SD = 13.1).

4.4. Stimuli

Participants saw scenarios where a single dinner party guest stayed idle at the table while the remaining guests worked on a collaborative task. In each scenario, we independently manipulated the number of dinner party guests (group size) and the amount of work to be done (workload). To ensure robustness, participants saw two classes of scenes. In Dish Clean-Up, agents were seated at the table with a number of plates in front of them. Participants were told that the agents needed to carry the plates to the sink, then wash and dry them. In Dish Clean-Up, workload was defined as the number of plates that need to be cleaned. In Salad Preparation, agents were also seated at the table with a number of carrots next to the sink. Participants were told that the agents needed to clean and chop the carrots to add them to the salad bowl. In Salad Preparation, workload was defined as the number of carrots that need to be prepared.

In each scene, the agents were depicted with unique colors. The red-colored agent was present in every scene in a fixed position, and this agent was always the loafer, or the guest that remained seated at the table without contributing to the task. Stimuli for this study were generated using Python's Matplotlib library. We represented a kitchen using a 7 × 7 grid, and programmatically populated all of the various layouts, objects, and agents in the scene. This approach allowed for a high degree of consistency across scenes.

We manipulated each scenario to reflect low (1 plate/carrot), medium (4 plates/carrots), or high (8 plates/carrots) workload, as well as small (2 agents), medium (4 agents), or large (6 agents) group size. This approach resulted in a total of 18 unique stimuli, or 3 workload levels × 3 group sizes × 2 kitchen scenarios, in a fully within-participants design. Participants were presented the 18 stimuli in random order. Representative stimuli are shown in Fig. 1.

4.5. Procedure

Participants were informed that every trial stimulus represented a different dinner party with new guests. In each trial, participants were told that the red agent remained seated at the table for the duration of the task. They were then asked to rate how acceptable they considered this guest's behavior. Participants reported their judgments on a Likert scale from 1 (completely unacceptable) to 10 (completely acceptable). Experiment 1 can be accessed at https://em.velezlab.opalstacked.com/Experiment1/.

4.6. Analysis

We analyzed our data using a linear mixed-effects model that predicted continuous judgments of idleness acceptability, $y_i \in [1, 10]$, as a function of the discrete fixed effects of workload x_1 and group size x_2. To enable pairwise comparisons between factor levels, workload and group size were each coded using two binary indicator variables measured against the baseline medium level (4 plates/carrots for workload, and 4 agents for group size). Thus, we use the notation $x_{1,1}$ to denote low workload, $x_{1,2}$ to denote high workload, $x_{2,1}$ to denote small group size, and $x_{2,2}$ to denote large group size. We also included all interactions between workload and group size levels, a fixed intercept term $\overline{\alpha}$, random intercepts across participants α_i, and residual error ϵ. These analyses were performed using Python's statsmodels, NumPy, and pandas.

4.7. Results

Results from Experiment 1 are presented in Fig. 3. There were significant main effects of workload and group size.

First, we found that participants judged idleness as more acceptable in larger groups. As group size increases, withholding effort is considered to be more acceptable. In particular, we found that participants judged idleness to be less acceptable in scenarios with smaller groups, compared to the medium-sized groups ($\beta_{2,1} = -1.01$, SE = 0.11, 95% CI = [−1.22, −0.80], $z = -9.41$, $p = 0.004$), and more acceptable in large groups, compared to medium-sized groups ($\beta_{2,2} = 0.57$, SE = 0.11, 95% CI = [0.36, 0.78], $z = 5.32$, $p < 0.001$).

Next, we found that participants judged idleness as less acceptable under high workloads. In particular, participants rated the idle agent's behavior as more acceptable in low workload scenarios, compared to the baseline of medium workload ($\beta_{1,1} = 1.00$, SE = 0.11, 95% CI = [0.79, 1.21], $z = 9.27$, $p < 0.001$), but less acceptable with high workloads compared to medium workloads ($\beta_{1,2} = -.31$, SE = 0.11, 95% CI = [−0.52, −0.10], $z = -2.89$, $p < 0.001$).

We found no significant interactions between workload and group size: (1) low workload and small group ($\beta = .12$, SE = 0.15, 95% CI = [−0.18, 0.42], $z = 0.78$, $p = 0.44$), (2) low workload and large group (β, SE = 0.15, 95% CI = [0.11, −0.54], $z = -1.61$, $p = 0.11$), high workload and small group (β, SE = 0.15, 95% CI = [−0.42, 0.18], $z = 0.43$, $p = 0.43$), (4) high workload and large group (β, SE = 0.15, 95% CI = [−0.18, 0.41], $z = 0.76$, $p = 0.45$).

We employed a simple model comparison approach to evaluate the explanatory power of different theoretical models and the necessity of our predictors. Importantly, our models make direct predictions without free parameters that need to be fit to data, allowing for strong tests of their explanatory power through both qualitative trend analysis and quantitative correlations with human judgments. The likelihood ratio test revealed that both workload (x_1) and group size (x_2) variables are statistically significant predictors of performance, with low p-values ($p < 0.001$) for all their coded variables ($x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2}$). These results indicate that removing either workload or group size variables would substantially diminish the model's explanatory power in predicting subject responses.

Table 1
Predictor significance in Experiment 1.

Predictor	LR	p-value
Low workload ($x_{1,1}$)	227.198	<0.001
High workload ($x_{1,2}$)	19.668	<0.001
Small group ($x_{2,1}$)	252.413	<0.001
Large group ($x_{2,2}$)	69.498	<0.001
Bottleneck (x_3)	N/A	N/A

Table 2
Model performance in Experiment 1.

Model	Spearman (r)	RMSE
Efficiency	0.828	1.509
Workload	0.632	2.160
Workload effort	0.158	3.266
SL linear	0.738	1.826
SL nonlinear	0.738	1.826
Norms	−0.738	4.690

In addition to these statistical tests, we computed the Spearman rank correlation and rank root mean squared error (RMSE) for each model as an indicator of fit to evaluate how well different theoretical models capture relative trends rather than absolute scales. Since these values are computed from a small number of means corresponding to the different conditions (a total of nine for Experiment 1), the differences across models are not statistically significant. Given these limitations, we report correlation and RMSE values descriptively. Overall, participants judged that it is less acceptable to withhold effort when there is more work to be done and fewer agents to do it. Notably, the social loafing model reflects the effect of group size, and the workload model reflects the effect of workload. Only the task efficiency model uniquely considers both of these variables within a single framework. Our task efficiency model demonstrated the highest correlation ($r = 0.83$) and lowest rank root mean square error ($RMSE = 1.51$) among the tested theoretical models. Full statistical details can be found in Tables 1 and 2.

However, some deviations between model predictions and human judgments remain, particularly in how people respond to differences in larger group sizes, which may be partly explained by participants' sensitivity to spatial factors like potential collisions between group members when there are many guests in the kitchen. These results provide evidence that participants evaluate idle collaborators based on their impact on task efficiency. Experiment 2 examines whether people additionally consider a factor that is uniquely motivated by the task efficiency model: the presence of environmental bottlenecks.

5. Experiment 2: Effects of resource bottlenecks

We further evaluate how people take into account task efficiency by manipulating the presence of serial bottlenecks that prevent division of labor. Our model of task speedup predicts that changing the number of bottlenecks in each scene – such as the number of sinks or cutting boards – should shift the saturation point at which adding more agents will no longer improve task completion. If there are a limited number of sinks to wash the dishes, or cutting boards to chop the carrots, then no matter how many agents are available, they cannot speed up the task.

5.1. Methods

As in Experiment 1, participants were again presented with dinner party scenarios where all agents except for one idle guest were actively engaged in a task. Participants were asked to evaluate the acceptability of this agent's idleness. The anonymized preregistration, which includes experimental design, data collection procedures, and our primary subject-level analysis, can be found at https://aspredicted.

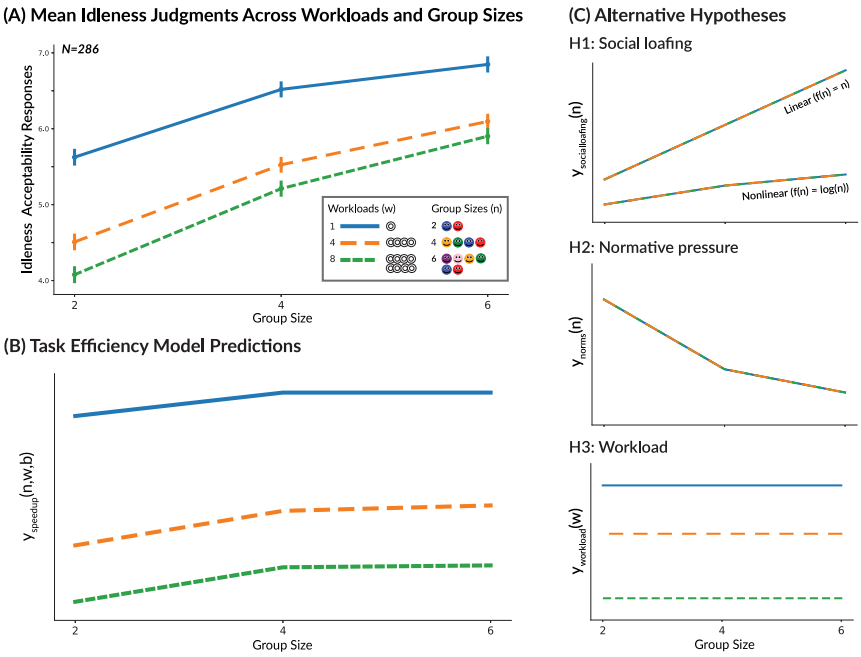

Fig. 3. Experiment 1 results. (A) Mean idleness judgments by workload and group size. Participants judged that it is more acceptable for the red agent to be idle in scenarios with larger groups and small workloads. Error bars represent standard errors. (B) Task efficiency model predictions, where idleness acceptability (y_{speedup}) is plotted by group size and workload. (C) Predictions from alternative hypotheses. H1–2 use simple heuristics to evaluate idleness based on group size, whereas H3 evaluates idleness based on workload. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

org/d2vq-vhc9.pdf. The experiment was performed with IRB approval (IRB # 15959); all participants provided informed consent prior to the experiment.

5.2. Participants

We recruited 300 English-speaking adults from the Prolific platform in exchange for monetary compensation ($2.60 for a 14 min experiment). Twenty-nine participants who failed an attention check were excluded from subsequent analysis, resulting in a total of 271 participants (82 male, 181 female, 3 non-binary; mean age = 39.0, SD = 12.9).

5.3. Stimuli

We designed stimuli that varied across three independent dimensions: (1) workload, (2) group size, and (3) bottlenecks, or the number of workstations available for agents to use. In Dish Clean-Up scenarios, workstations were defined as the number of sinks. In Salad Preparation scenarios, workstations were defined as the number of cutting boards.

Stimuli for Experiment 2 were generated using the same methods as Experiment 1. In each stimulus, every agent was depicted with a unique color. The red-colored agent was present in every scene in a fixed position, and this agent was always the idle guest that remained seated at the table without contributing to the task.

As in Experiment 1, we independently varied group size (1, 4, or 6 agents) and workload (1, 4, or 8 units of work). We additionally manipulated the presence of environmental bottlenecks: high-bottleneck scenarios had only 1 workstation, while low-bottleneck scenarios had 6 workstations. This approach resulted in a total of 36 unique stimuli, or 3 workload levels × 3 group sizes × 2 bottleneck levels × 2 kitchen scenarios, in a fully within-participants design. Stimuli were presented in random order. Representative stimuli are shown in Fig. 1.

5.4. Procedure

Participants followed the same procedure outlined in Experiment 1. Experiment 2 can be accessed at https://em.velezlab.opalstacked.com/Experiment2/.

5.5. Analysis

Analyses were performed using the same methods as Experiment 1, with the addition of a two-level discrete fixed effect for bottleneck, x_3, as well as all possible two-way interactions between workload, x_1, group size, x_2, and bottleneck. To analyze these effects, we employed a linear mixed-effects model where the continuous output criterion was a judgment of idleness acceptability, $y_i \in [1, 10]$.

5.6. Results

Results from Experiment 2 are presented in Fig. 4. There were significant main effects of workload, group size, and bottlenecks. First, we replicated the effects of workload and group size produced in Experiment 1. Smaller group sizes produced a significant decrease in idleness acceptability compared to the medium-sized groups ($\beta_{2,1}$ = −1.30, SE = 0.10, 95% CI = [−1.50, −1.11], z = −13.06, $p < 0.001$), whereas larger group sizes than the medium baseline led to a significant increase in idleness acceptability ($\beta_{2,2}$ = 0.624, SE = 0.10, 95% CI = [0.43, 0.82], z = 6.26, $p < 0.001$). Low workload produced a significant increase in idleness acceptability compared to the baseline of medium workload ($\beta_{1,1}$ = 1.43, SE = 0.10, 95% CI = [1.23, 1.63], z = 14.36, $p < 0.001$), whereas high workloads compared to the medium baseline produced a significant decrease in idleness acceptability ($\beta_{1,2}$ = −0.23, SE = 0.10, 95% CI = [−0.43, −0.04], z = −2.31, p = 0.02). Together, these results provide further evidence that, as workload increases and group size decreases, idleness is judged to be less acceptable.

Next, we found that participants' judgments were also sensitive to the number of workstations available. In particular, participants judged idleness to be less acceptable when there were more workstations available (β_3 = −0.73, SE = 0.09, 95% CI = [−0.91, −0.55], z = −8.05, $p < 0.001$).

Some of the estimates for the fixed effects of the interactions between workload, group size, and bottlenecks were significant. Specifically, significant interactions were found in the following cases, which qualitatively fit predictions from our task efficiency model: (1) low workload and large group (β = −0.33, SE = 0.12, 95% CI = [−0.57,

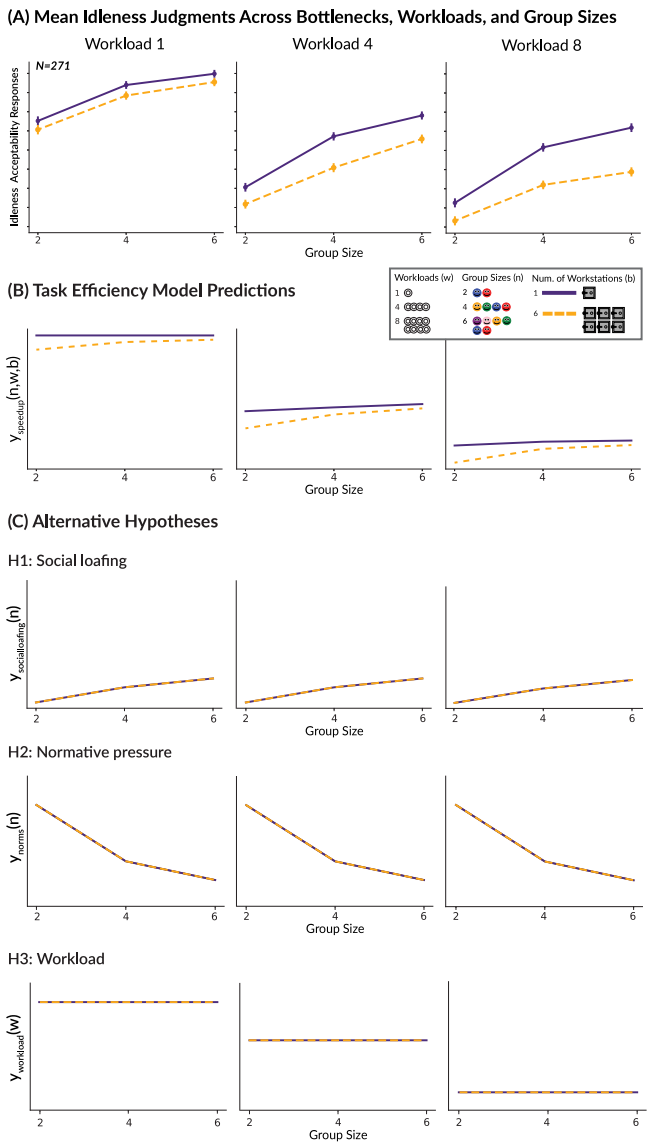

Fig. 4. Experiment 2 results. (A) Mean idleness judgments by group size, workload, and bottlenecks. Participants' idleness judgments align with results from Experiment 1, and additionally show that idleness is more acceptable with fewer workstations. Error bars represent standard errors. (B) Task efficiency model predictions, where idleness acceptability ($y_{speedup}$) is plotted for the same group sizes, workloads, and bottlenecks. (C) Hypotheses from alternative models, which do not take into account environmental bottlenecks when evaluating idleness.

−0.09], $z = -2.70$, $p = 0.007$). The effect of group size on idleness judgments was reduced in larger groups, where there were more group members than needed for the workload. The interaction between (2) small group and bottleneck was also significant ($\beta = 0.31$, SE = 0.10, 95% CI = [0.11, 0.50], $z = 3.09$, $p = 0.002$). As shown in Fig. 4, the effect of the number of workstations on acceptability judgments was reduced in smaller groups, where there are not enough guests to fill the workstations. Finally, there were significant interactions between (3) low workload and bottleneck ($\beta = 0.38$, SE = 0.10, 95% CI = [0.18, 0.57], $z = 3.78$, $p < 0.001$), and (4) high workload and bottleneck ($\beta = -0.252$, SE = 0.10, 95% CI = [−0.45, −0.06], $z = -2.53$, $p = 0.01$). That is, the number of workstations had a greater effect on idleness judgments with high workloads compared to low workloads. Intuitively, this finding makes sense. When there is only one dish to clean, or one carrot to chop, additional workstations should have little effect on task completion. However, with more work to be done, more guests can speed up task completion if more workstations are available.

Table 3
Model performance in Experiment 2.

Model	Spearman (r)	RMSE
Efficiency	0.950	0.816
Workload	0.843	1.414
Workload effort	0.264	3.055
SL linear	0.474	2.582
SL nonlinear	0.474	2.582
Norms	−0.474	4.320

Table 4
Predictor significance in Experiment 2.

Predictor	LR	p-value
Low workload ($x_{1,1}$)	5.996	0.199
High workload ($x_{1,2}$)	12.950	0.012
Small group ($x_{2,1}$)	12.101	0.017
Large group ($x_{2,2}$)	2.163	0.706
Bottleneck (x_3)	11.668	0.040

None of the remaining interaction effects were significant: (5) low workload and small group ($\beta = 0.23$, SE = 0.12, 95% CI = [−0.01, 0.46], $z = 1.85$, $p = 0.06$), (6) high workload and small group ($\beta = -0.05$, SE = 0.12, 95% CI = [−0.29, 0.19], $z = -0.40$, $p = 0.69$), (7) high workload and large group ($\beta = -0.22$, SE = 0.12, 95% CI = [−0.46, 0.02], $z = -1.79$, $p = 0.07$), (8) large group and bottleneck ($\beta = 0.03$, SE = 0.10, 95% CI = [−0.16, 0.23], $z = 0.33$, $p = 0.74$).

Similar to our model comparisons for Experiment 1, we evaluated the explanatory power of each theoretical model and the necessity of our predictors. The likelihood ratio test revealed that bottlenecks were statistically significant, with a p-value of 0.040 for coded variable x_3. Our analysis revealed important interactions between bottlenecks, workload, and group size. When workloads were high, acceptability was primarily determined by the presence or absence of bottlenecks. In contrast, for small groups, performance was predominantly influenced by the workload level. These results indicate that removing workload, group size, or bottlenecks would substantially diminish explanatory power. In addition to these statistical tests, we computed the correlation and RMSE for each model as an indicator of fit. Since these values are computed from a small number of means corresponding to the different conditions (a total of 18 for Experiment 2), the differences across models are not statistically significant. Our task efficiency model demonstrated the highest correlation ($r = 0.853$) and lowest root mean square error ($RMSE = 2.809$), best capturing trends in subject responses. Full statistical details can be found in Tables 3 and 4.

Serial bottlenecks prevent division of labor, even with more available collaborators and tasks to be completed. These environmental constraints, such as a limited number of tools or workstations, inherently limit the marginal impact an additional group member can have, regardless of how much work is left. Therefore, if people consider task efficiency when judging idleness, they should base their judgments not only on workload and group size (as in Experiment 1), but also on bottlenecks that inhibit parallelization. In Experiment 2, we found that both workload and group size were significant predictors of idleness acceptability, replicating our results from Experiment 1. Additionally, the number of workstations predicted participants' judgments. Participants judged idleness to be less acceptable with more workstations — and thus fewer bottlenecks to parallelization — across varying workload levels and group sizes. Lastly, we found interactions between bottlenecks, workload, and group size that are qualitatively captured by the task efficiency model. Intuitively, participants' judgments reflect that the benefit of adding workstations is limited by workload and group size.

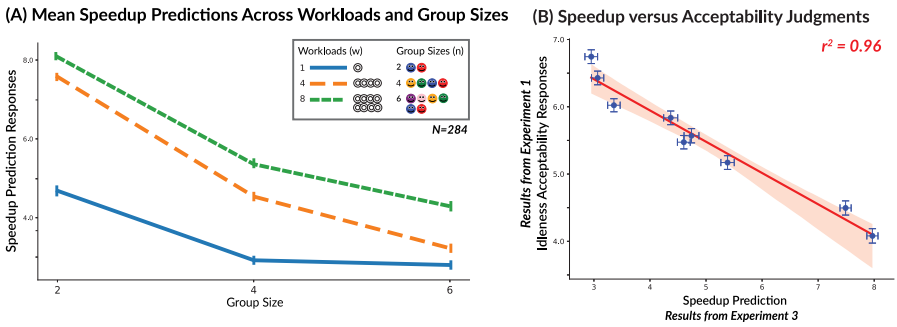

Fig. 5. Experiment 3 results. (A) Mean speedup predictions by workload and group size. Participants predicted that an idle collaborator would have a greater impact on a task being performed faster in scenarios with smaller groups and larger workloads. Error bars represent standard error. (B) Inverse relationship between idleness acceptability and speedup. Participant judgments of idleness were inversely correlated with predictions of task speedup ($r = -0.98, p < 0.001$), indicating that people judge idleness more harshly when the idle guest could have sped up the task more.

6. Experiment 3: Inverse relationship between speedup predictions and idleness acceptability

In the previous sections, we found significant effects of the variables that impact task parallelizability — group size, workload, and resource bottlenecks — on people's judgments of idleness during a collaborative task. However, up until this point we have assumed that speedup is the latent factor driving these judgments. In our third experiment, we directly test this assumption by explicitly asking subjects to estimate how much faster they believe the task would have been completed if the idle collaborator had contributed.

6.1. Methods

As in Experiments 1 and 2, participants were presented with dinner party scenarios where all agents except for one idle guest were actively engaged in a task. Participants were asked to evaluate how much faster the task would have been completed if the idle guest had helped. The anonymized preregistration, which includes experimental design, data collection procedures, and our primary subject-level analysis, can be found at https://aspredicted.org/34r2-573j.pdf. The experiment was performed with IRB approval (IRB # 15959); all participants provided informed consent prior to the experiment.

6.2. Participants

We recruited 300 English-speaking adults from the Prolific platform in exchange for monetary compensation ($1.20 for a 6 min experiment). Sixteen participants who failed an attention check were excluded from subsequent analysis, resulting in a total of 284 participants (132 male, 145 female; mean age = 41.4, SD = 14.2).

6.3. Stimuli

To maintain consistency with previous experiments, we used the same set of 18 stimuli from Experiment 1 that varied along two independent dimensions across two kitchen scenarios: (1) workload (1, 4, or 8 units of work), and (2) group size (1, 2, or 6 agents).

6.4. Procedure

Participants were informed that every trial stimulus represented a different dinner party with new guests. In each trial, participants were told that the red agent remained seated at the table for the duration of the task. They were then asked to rate how much faster the task would have been completed if the red guest had contributed. Participants reported their judgments on a Likert scale from 1 (not faster) to 10 (much faster). Experiment 3 can be accessed at https://em.velezlab.opalstacked.com/Experiment3/.

6.5. Analysis

We analyzed our data using a linear mixed-effects model that predicted continuous judgments of task speedup, $y_i \in [1, 10]$, as a function of the discrete fixed effects of workload x_1 and group size x_2. Similar to the analysis in Experiment 1, workload and group size were each coded using two binary indicator variables measured against the baseline medium level (4 plates/carrots for workload, and 4 agents for group size). We also included all interactions between workload and group size levels, a fixed intercept term, random intercepts across participants, and residual error. These analyses were performed using Python's statsmodels, NumPy, and pandas packages.

6.6. Results

Results from Experiment 3 are presented in Fig. 5. There were significant main effects of workload and group size, as well as a highly negative correlation between speedup predictions in Experiment 3 and idleness acceptability judgments in Experiment 1.

First, we found that participants predicted greater speedup if the idle guest participated in smaller groups. As group size increases, the potential speedup that could be gained from an additional agent participating decreases. In particular, we found that participants judged speedup to be fastest with smaller groups, compared to medium groups ($\beta_{2,1} = 3.042$, SE = 0.14, 95% CI = [2.77, 3.32], $z = 21.84$, $p < 0.001$), and slowest in large groups compared to medium-sized groups ($\beta_{2,2} = -1.30$, SE = 0.14, 95% CI = [−1.58, −1.03], $z = -9.35$, $p < 0.001$). Next, we found that participants predicted that an idle agent would speed up a task most given high workloads. In particular, participants judged potential speedup to be greatest in high workload scenarios compared to the baseline medium workload ($\beta_{1,2} = 0.82$, SE = 0.14, 95% CI = [0.55, 1.01], $z = 5.92$, $p < 0.001$), and lowest in low workload scenarios compared to the baseline medium workload ($\beta_{1,1} = -1.61$, SE = 0.14, 95% CI = [−1.88, −1.34], $z = -11.55$, $p < 0.001$).

Some of the interactions between workload and group size had significant effects on speedup predictions, specifically in scenarios with the lowest workload level. The effect of low workload was reduced in smaller groups ($\beta = -1.28$, SE = 0.20, 95% CI = [−1.67, −0.90], $z = -6.52$, $p < 0.001$) and increased in larger groups ($\beta = 1.18$, SE = 0.20, 95% CI = [0.79, 1.56], $z = 5.96$, $p < 0.001$), which reflects the previous finding in Experiment 2 that judgments of speedup and idleness change when there are more guests than the amount of work needed to be done. There were no significant interactions between high workload and small groups ($\beta = -0.34$, SE = 0.20, 95% CI = [−0.72, 0.05], $z = -1.71$, $p = 0.09$) or between high workload and large groups ($\beta = 0.23$, SE = 0.20, 95% CI = [−0.16, 0.62], $z = 1.17$, $p = 0.24$).

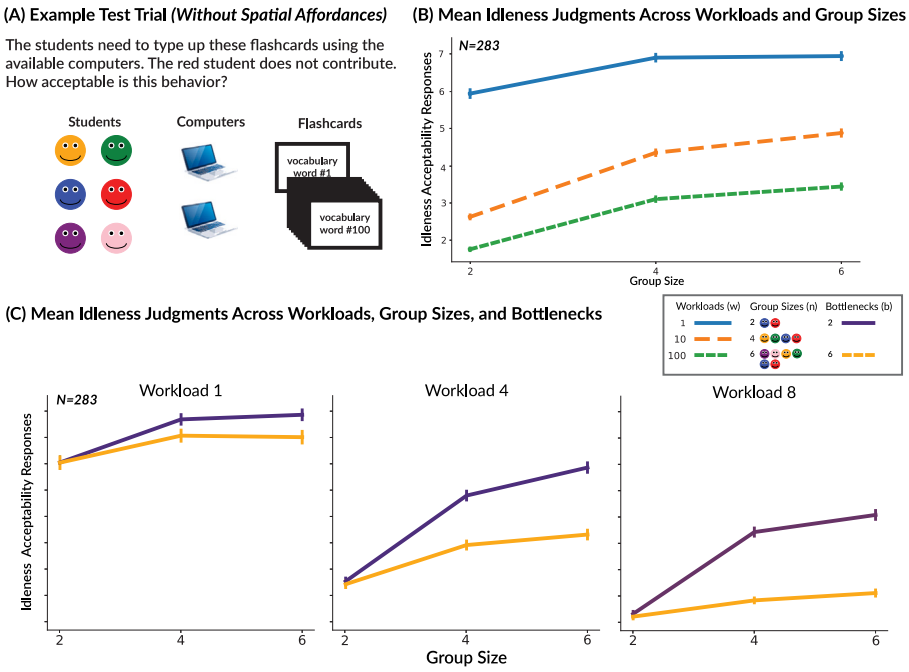

Fig. 6. Experiment 4 results. (A) An example trial from the experiment. Participants were shown different groups of students, numbers of computers (bottlenecks), and flashcards (workload), and asked to evaluate an idle collaborator's behavior. (B) Mean idleness judgments by workload and group size. Participants predicted that it would be less acceptable for a collaborator to remain idle in scenarios with smaller groups and larger workloads. (C) Mean idleness judgments by workload, group size, and bottleneck. Participants predicted that it would be more acceptable for a collaborator to remain idle in scenarios with more available tools. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6.7. Speedup predictions versus idleness acceptability

Finally, to test whether participants' judgments of idleness were related to their perceptions of potential task speedup if the idle agent contributed, we computed a Pearson correlation between mean ratings of speedup in Experiment 3 and idleness acceptability in Experiment 1 aggregated by workload and group size. The analysis revealed a highly significant negative correlation ($r = -0.98, p < 0.001$), suggesting that conditions in which participants perceived the greatest potential speedup from an additional collaborator corresponded to idleness being less acceptable. This result supports our hypothesis that participants' judgments of acceptability are systematically influenced by task efficiency, even when speedup is not explicitly considered.

7. Experiment 4: Removing spatial affordances

There is a key factor in the kitchen scenes that is not accounted for by our model's predictions: spatial affordances. Larger groups face a more difficult coordination problem because there is a greater chance of colliding with another guest, thus complicating task completion. Indeed, we see deviations between model predictions and participants' judgments in Experiment 1; with larger groups, participants predict an increase in the acceptability of remaining idle, even when the number of available tools should make the ability for an additional guest to contribute the same. In the final experiment, we remove this confound by asking subjects to evaluate a collaborative scenario without spatial navigation, in which students must type up vocabulary words using a set number of available computers. This design highlights how structural constraints on task decomposability and resource availability influence collaborative judgments, even in the absence of spatial interference. As a first step, it provides a controlled contrast to help explain when and why the original task efficiency model deviates from human judgments. Importantly, the scenario mirrors everyday collaborative contexts, such as shared office work or group projects, where spatial constraints may be minimal but coordination challenges persist.

7.1. Methods

Participants were presented with group project scenarios where all students except for one were actively engaged in the task. Participants were then asked to evaluate how acceptable this idle student's behavior was. The anonymized preregistration, which includes experimental design, data collection procedures, and our primary subject-level analysis, can be found at https://aspredicted.org/4669-44ng.pdf. The experiment was performed with IRB approval (IRB # 15959); all participants provided informed consent prior to the experiment.

7.2. Participants

We recruited 300 English-speaking adults from the Prolific platform in exchange for monetary compensation ($1.00 for a 5 min experiment). Seventeen participants who failed an attention check were excluded from subsequent analysis, resulting in a total of 283 participants (121 male, 154 female, 4 non-binary; mean age = 38.9, SD = 12.7).

7.3. Stimuli

Participants were shown scenarios where a single student stayed idle while their remaining group members typed up vocabulary words using computers with flashcard software. In each scenario, we independently manipulated the number of students (group size), the amount of work to be done (flashcards), and bottlenecks (computers). In each scene, the red-colored agent was the idle student. The students, computers, and flashcards were presented in three groups without indicating spatial positions. We manipulated each scenario to reflect low (1 flashcard), medium (10 flashcards), or high (100 flashcards) workload, small (2 students), medium (4 students), or large (6 students) group size, and small (2 computers) or large (6 computers) numbers of tools. This approach resulted in a total of 18 unique stimuli, or 3 workload levels × 3 group sizes × 2 bottleneck levels, in a fully within-participants design. Participants were presented the 18 stimuli in random order. Representative stimuli are shown in Fig. 6A.

7.4. Procedure

Participants were informed that every trial stimulus represented a different group of students. In each trial, participants were that the students need to type up these flashcards using the available computers. The red student does not contribute. They were then asked to rate how acceptable this behavior is on a Likert scale from 1 (completely unacceptable) to 10 (completely acceptable). Experiment 4 can be accessed at https://em.velezlab.opalstacked.com/Experiment4/.

7.5. Analysis

Analyses were performed using the same methods as Experiment 2. To analyze these effects, we employed a linear mixed-effects model where the continuous output criterion was a judgment of idleness acceptability, $y_i \in [1, 10]$.

7.6. Results

Results from Experiment 4 are presented in Fig. 6. There were significant main effects of workload, group size, and bottlenecks, replicating the results of Experiments 1 and 2. Smaller group sizes produced a significant decrease in idleness acceptability compared to the medium-sized groups ($\beta_{2,1} = -2.41$, SE = 0.16, 95% CI = [−2.72, −2.11], $z = -15.33$, $p < 0.001$), whereas larger group sizes than the medium baseline led to a significant increase in idleness acceptability ($\beta_{2,2} = 0.72$, SE = 0.16, 95% CI = [0.41, 1.02], $z = 4.53$, $p < 0.001$). Low workload produced a significant increase in idleness acceptability compared to the baseline of medium workload ($\beta_{1,1} = 2.27$, SE = 0.16, 95% CI = [1.96, 2.58], $z = 14.44$, $p < 0.001$), whereas high workloads compared to the medium baseline produced a significant decrease in idleness acceptability ($\beta_{1,2} = -1.16$, SE = 0.16, 95% CI = [−1.47, −0.85], $z = -7.36$, $p < 0.001$). We found that participants' judgments were also sensitive to the number of workstations available. In particular, participants judged idleness to be less acceptable when there were more workstations available ($\beta_3 = -1.45$, SE = 0.14, 95% CI = [−1.73, −1.17], $z = -10.09$, $p < 0.001$). Together, these results provide further evidence that, as workload increases and group size or number of available tools decreases, idleness is judged to be less acceptable.

Some of the estimates for the fixed effects of the interactions between workload, group size, and bottlenecks were significant. Specifically, significant interactions were found in the following cases, which qualitatively fit predictions from our task efficiency model: (1) low workload and small group ($\beta = 0.80$, SE = 0.19, 95% CI = [0.42, 1.18], $z = 4.14$, $p < 0.001$). The acceptability of idleness tended to increase with low workloads in small groups, since there was only one flashcard to make. The interaction between (2) low workload and large group was also significant ($\beta = -0.51$, SE = 0.19, 95% CI = [−0.89, −0.13], $z = -2.65$, $p = 0.008$). The effect of group size on idleness judgments was reduced in larger groups, where there were more group members than needed for the workload. The interaction between (3) small group and bottleneck was also significant ($\beta = 1.23$, SE = 0.16, 95% CI = [0.92, 1.54], $z = 7.78$, $p < 0.001$). The effect of the number of computers on acceptability judgments was reduced in smaller groups, where there are not enough guests to fill the workstations. There were also significant interactions between (4) low workload and bottleneck ($\beta = 0.78$, SE = 0.16, 95% CI = [0.47, 1.09], $z = 4.95$, $p < 0.001$). Again, the number of workstations had a lower effect on idleness judgments with low workloads, since additional computers should have little impact if there is only one flashcard to create.

None of the remaining interaction effects were significant: (5) high workload and bottleneck ($\beta = -0.27$, SE = 0.16, 95% CI = [−0.58, 0.03], $z = -1.74$, $p = 0.08$), (6) high workload and large group ($\beta = -0.20$, SE = 0.19, 95% CI = [−0.58, 0.18], $z = -1.02$, $p = 0.31$), (7) large group and bottleneck ($\beta = -0.32$, SE = 0.16, 95% CI = [−0.63, 0.00], $z = -2.02$, $p = 0.044$), and (8) high workload and small group ($\beta = 0.39$, SE = 0.19, 95% CI = [0.01, 0.77], $z = 2.02$, $p = 0.043$).

Table 5
Model performance in Experiment 4.

Model	Spearman (r)	RMSE
Efficiency	0.983	0.471
Workload	0.949	0.816
Workload effort	0.949	0.816
SL linear	0.246	3.055
SL nonlinear	0.246	3.055
Norms	−0.246	4.000

Table 6
Predictor significance in Experiment 4.

Predictor	LR	p-value
Low workload ($x_{1,1}$)	1139.804	<0.001
High workload ($x_{1,2}$)	241.558	<0.001
Small group ($x_{2,1}$)	372.071	<0.001
Large group ($x_{2,2}$)	19.271	<0.001
Bottleneck (x_3)	360.180	<0.001

Experiment 4 eliminated the spatial confounds present in previous kitchen-based scenarios by examining judgments about students typing vocabulary words with limited computers. The results showed significant main effects of workload, group size, and bottlenecks, strongly aligning with model predictions and providing an even closer match to model predictions than Experiment 1. Smaller groups and higher workloads decreased the acceptability of idleness, while larger groups and lower workloads increased it. The availability of workstations also influenced judgments, with more workstations leading to lower acceptability of idleness.

Similar to our model comparisons for Experiments 1 and 2, we evaluated the explanatory power of each theoretical model and the necessity of our predictors. The likelihood ratio test revealed that all three variables in the task efficiency model were statistically significant predictors of subject responses ($p < 0.001$). Additionally, our task efficiency model demonstrated the highest correlation ($r = 0.983$) and lowest root mean square error ($RMSE = 0.471$). Full statistical details can be found in Tables 5 and 6.

7.7. Modeling coordination costs

Given that the first three experiments present agents in spatial environments, and in Experiment 4 agents perform a cognitive task, can the task efficiency model predict differences in "coordination costs" that people attribute to different kinds of collaborations? For example, larger teams may face additional coordination costs in spatial environments in order to prevent colliding with one another and blocking the space. Collaborations may additionally involve ambiguity about who should perform which subtask, contributing to coordination overhead that is not captured by the idealized assumptions of Amdahl's Law (Dessein & Santos, 2006). To explore potential deviations between the task efficiency model and participants' idleness judgments across environments with differing coordination complexity, we conducted an exploratory analysis to augment the model with a *coordination cost coefficient* α, such that $c = \alpha \cdot (N - 1)$. This parameter scales with larger group sizes, capturing the more complex coordination challenges that arise as more agents are coordinating their behaviors. It assumes that coordination overhead grows linearly with the number of team members, preserving the interpretability and structure of Amdahl's Law while accounting for realistic deviations due to constraints on team performance. Specifically, this coefficient constrains the speed-up for each subtask via:

$$s(n, b_i, \alpha) = \frac{\min(n, b_i)}{c \cdot \min(n - 1, b_i)} \tag{9}$$

To evaluate model fit, we compared the baseline task efficiency model (assuming no coordination cost) to the model including α for Experiments 1 and 4. Both models used shared linear rescaling parameters (a, b) to account for differences in scale between the task efficiency model and subject responses. The coordination cost model fit the data substantially better: AIC decreased by 278.23 and BIC by 263.77 relative to the baseline, indicating improvement in explanatory power.

These results confirm our hypothesis that participants implicitly incorporate coordination constraints when evaluating team efficiency. In Experiment 1, group members were faced with additional coordination challenges in order to move between the table, counters, and sinks. Indeed, when fit to participants' judgments using mean-squared error, α was substantially larger for Experiment 1 ($\alpha_1 = 8.88$) than 4 ($\alpha_4 = 4.84$). These results demonstrate that people attribute higher coordination costs to spatial environments, which require division of labor as well as coordination in physical space. This analysis supports the idea that people evaluate idle collaborators with respect to both their counterfactual impact on team throughput and the practical coordination costs incurred by adding more agents. The coordination cost parameter offers a principled way to quantify and compare collaboration friction across environments, further strengthening the generality of our findings.

8. Discussion

Idleness is often perceived as group members selfishly withholding effort during collaborations, thereby reducing group productivity. However, research in distributed computing systems offers an alternative interpretation. Namely, adding collaborators to a task is not guaranteed to speed up task completion, but depends on workload, current contributors, and serial bottlenecks that limit parallelization. Thus, individuals may remain idle because their efforts will not help the group complete tasks more efficiently. Here, we assess whether people consider task efficiency when evaluating idle collaborators, and we present a modeling framework inspired by distributed systems to formalize factors that likely influence structural constraints on collaborative performance.

We derived a task efficiency model, inspired by Amdahl's Law, that evaluates whether it is acceptable to remain idle based on how much the idle collaborator's contribution would have sped up task completion. Our findings reveal a qualitative correspondence between model predictions and participant ratings of idleness acceptability. The model predicts that withholding effort is less acceptable in situations with high workloads, smaller groups, and fewer bottlenecks to prevent division of labor. We validated these predictions with idleness judgments collected from 1124 participants across various scenarios (see Fig. A.8). In Experiment 1, the model accurately captured expected trends for workload and group size, including the plateau in medium-sized groups due to the limited number of sinks and cutting boards. Experiment 2 independently manipulated resource bottlenecks, revealing that participants judged idleness to be less acceptable when more workstations were available—an effect also predicted by the model. Experiment 3 provided direct evidence linking idleness judgments to perceptions of task efficiency by showing a strong correlation between participants' estimates of potential task speedup and their evaluations of idleness. Experiment 4 further validated this relationship by replicating these effects in a scenario that removed spatial affordances and contextual norms associated with dinner parties, demonstrating the generalizability of our findings. Together, these findings demonstrate that people evaluate idle collaborators based on the task speedup that would have occurred if they had contributed. We also considered several baseline models that take into account workload, typical social loafing behavior, and normative social pressures. None of these models presented a qualitative fit to the trends found in human responses.

Our results suggest that people evaluate idleness using counterfactual reasoning, specifically by considering how group members could have acted differently. This aligns with prior work on Counterfactual Simulation Models (Gerstenberg et al., 2023, 2018; Wu & Gerstenberg, 2024; Wu et al., 2023; Xiang, Landy et al., 2023). Our contribution is to provide a domain-specific instantiation of counterfactual reasoning grounded in principles from distributed systems. Although our model does not simulate rich causal mechanisms, it allows us to reason about a specific class of counterfactuals relevant to team efficiency: How would task completion time change if an agent had contributed differently, given current task constraints? By fixing structural features like bottlenecks and workload, we quantify how much an individual's absence would delay group success. In this sense, distributed systems theory helps formalize which counterfactuals are most relevant in collaborative environments—those tied to functional interdependencies and performance bottlenecks. This work complements existing counterfactual models by demonstrating how focusing on structured environments can yield precise, testable predictions about individual contributions to collective outcomes. However, our current approach models counterfactual impact under the simplifying assumption that the actions of other agents remain fixed. Recent work on counterfactual reasoning suggests that people may adopt more flexible, probabilistic perspectives, where outcomes are uncertain and multiple alternatives are considered (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012; Quillien & Lucas, 2023). By this account, an idle collaborator may be judged not just for what would have happened if they had acted, but for how their presence could have made success more likely by increasing the robustness of the group's performance in the face of potential failures. This reflects a broader intuition that collaboration is not just about maximizing throughput, but also about reducing variance and guarding against risk. Future work should extend our current framework by modeling probabilistic group outcomes given that some collaborators may fail, underperform, or refuse to cooperate, drawing on ideas from fault tolerance and reliability in distributed systems to capture the value of 'backup' agents (Cristian, 1991).

Our results provide initial evidence that theories from distributed systems can provide a source of hypotheses about how humans navigate the structural constraints that tasks and environments place on collaborations (Vélez et al., 2023). By isolating these constraints, we gain a principled way to predict how changes in task structure influence collective efficiency. This abstraction allows us to pinpoint when social or cognitive factors must be invoked to explain deviations from structural predictions. Thus, it is also important to consider what aspects these models do not capture. Human groups are subject to unique constraints that are not fully comparable to distributed computing systems, including ambiguities in task allocation, limitations on communication, differences in competence, power dynamics, social inference, fairness, and fatigue (Baer & Odic, 2022; Hawkins et al., 2023; Marjieh, Gokhale, Bullo, & Griffiths, 2024; Messeri, Bicchi, Zanchettin, & Rocco, 2022; Xiang, Landy et al., 2023; Yu & Thompson, 2024). Each of these constraints gives rise to dynamics that are distinctive to human collaborations. For instance, participants were sensitive to group size even when only one unit of work was present. In distributed systems, the number of agents should not affect the time to complete an isolated task, yet idleness was judged more acceptable in larger groups. This result may reflect ambiguities in task allocation, since unlike scheduling algorithms, human groups often lack clarity on task assignments without explicit communication (Brennan, Chen, Dickinson, Neider, & Zelinsky, 2008; Davis & Burns, 2011). It is possible that participants' judgments take into account what will happen if none of the guests stand up to complete the task, which is more probable when there are only two agents present (Kwon, Zhi-Xuan, Tenenbaum, & Levine, 2023). Another key distinction between human groups and distributed systems is heterogeneity in individual capabilities. Unlike identical computing nodes, people differ in skills, experience, and task proficiency, which shape both actual performance and perceived contributions (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). For example, expectations around acceptable idleness may vary between children learning to cook versus novices unfamiliar with the task.

One promising approach to modeling this variation is heterogeneous job scheduling algorithms, which account for performance differences across machines (Braun et al., 2001). Future work may explore to what extent these algorithms capture how humans assess the contributions of collaborators with varying skill levels.

Unlike distributed systems, participants' judgments are also certainly influenced by social concerns, causing divergences from model predictions. It may be counter-normative to avoid helping altogether, even when doing so would have no impact on speedup. In Experiment 2, participants consistently overestimated the benefit of having more workstations, suggesting that the presence of additional sinks increases social pressure to contribute despite diminishing returns. Idleness was judged especially harshly when the number of helpers exceeded the number of sinks, such as when five agents were cleaning but only one sink was available. In such cases, participants may infer that group members are expected to signal helpfulness, regardless of actual utility (Legare & Wen, 2014). These discrepancies between model predictions and subject responses, especially with higher workloads, larger group sizes, and fewer tools, may also be accounted for by fairness considerations, such as the expectation of turn-taking with higher workloads (Tabibnia & Lieberman, 2007; Xiang, Vélez and Gershman, 2023). Culturally-specific norms may also influence idleness judgments, such as whether an invited guest should contribute to housework (Cialdini & Goldstein, 2004; Earley, 1989). Despite these limitations, distributed systems theory supplements existing models of human collaboration by giving us heuristics to evaluate structural constraints that environments, tasks, and group topologies place on the ability to distribute labor and resources. Distributed systems provide a theoretical upper bound on team performance, while models of human cognition explain deviations from this upper bound by considering human goals and rich psychological processes. We can begin to think of human groups as natural distributed systems, encompassing some limited aspects of distributed computing while also including unique properties of human cognition.

Our work provides a starting point to examine collaborative behavior from a distributed systems perspective. Moving forward, this work can be extended in at least three ways. First, our model makes assumptions about the given tasks and environments that impose limitations to its ecological validity. Because we designed stimuli that are static screenshots of time-dependent tasks, we must approximate the relative proportions of overall task time required to complete each subtask. We should confirm the model's validity and replicate these findings in dynamic multi-player simulations (Shrout & Rodgers, 2018; Yarkoni, 2022). Second, our framework establishes a computational upper bound on division of labor that can be used to systematically study active collaborative behavior. Given that people's judgments align with theoretical constraints on group efficiency, we can now investigate how this understanding manifests in real-time collaborations. Our model makes quantitative predictions that could be tested in dynamic multi-player paradigms where participants allocate roles, resources, and subtasks in real time. For example, in environments like Overcooked (Carroll et al., 2019), we could examine how performance compares to model predictions under different conditions. Groups might deviate from optimal behavior when faced with different kinds of tasks, skills, and scenarios. Third, our framework can be extended to capture properties that are *unique* to human collaborations, such as how considerations of task efficiency may interact with culturally-specific norms, as well as which aspects of collaboration are shared with other species. Although explicit punishment of non-cooperators does not apply to primates (Riedl, Jensen, Call, & Tomasello, 2012), research suggests that some species may regulate cooperation through fairness-like processes such as disadvantageous inequity aversion (Brosnan, 2011; McAuliffe & Thornton, 2015). Our task efficiency framework provides a quantitative foundation for exploring these evolutionary questions by establishing a baseline against which both human and non-human collaborative behaviors can be systematically compared and contrasted.

Collaboration enables humans to achieve goals that no one individual could do on their own. However, finding ways to efficiently combine our efforts is itself a challenge. Here, we propose that distributed systems provide a rich source of hypotheses about how the structure of collaborative groups impacts performance, and we find that theories inspired by these systems capture human judgments about when it is acceptable not to contribute to a collaborative task. Put together, our work provides the first steps towards understanding human collaborations as a natural distributed system.

CRediT authorship contribution statement

Elizabeth Mieczkowski: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Cameron Turner:** Validation, Methodology, Formal analysis. **Natalia Vélez:** Writing – review & editing, Supervision, Conceptualization. **Thomas L. Griffiths:** Writing – review & editing, Supervision, Conceptualization.

Resources for reproducibility

The full stimulus set, behavioral dataset, and codebase can be found at https://github.com/emieczkowski/IdleCollaborators.

Acknowledgment

This work was supported in part by the Department of Defense (DoD) through the National Defense Science and Engineering Graduate (NDSEG) Fellowship Program and a grant from the Templeton World Charity Foundation.

Appendix. Supplemental materials

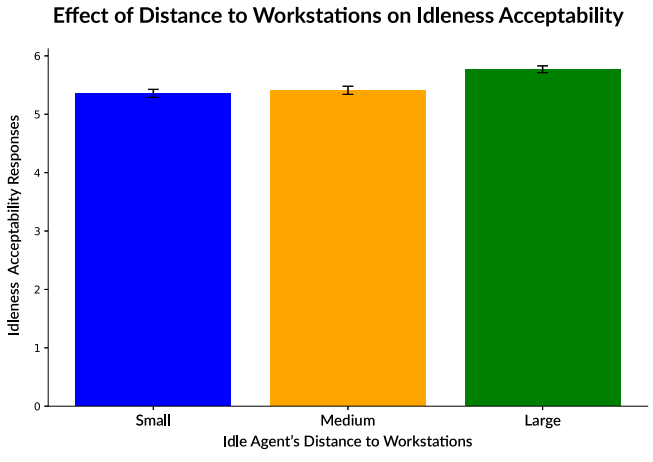

Fig. A.7. Experiment 5 Results. The distance of the idle agent from the workstations needed to complete the task was varied across workloads (1, 4, 8) and group sizes (2, 4, 6) in both Dish Clean-Up and Salad Preparation. There was a small but significant effect when the agent was farther from the workstations, but not when the agent was closer, versus the medium distance condition. This finding suggests that subjects take spatial information and potential effort into account when evaluating idleness, but that these effects are small compared to the effects of workload, group size, and bottlenecks.

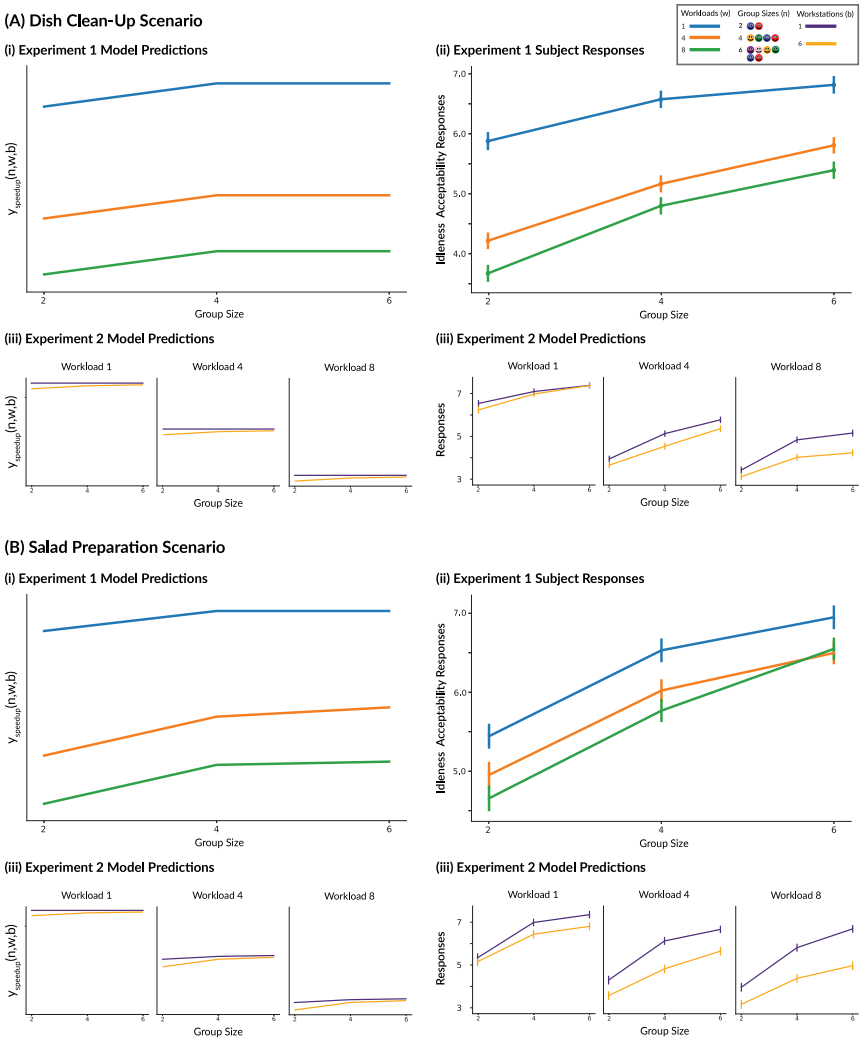

Fig. A.8. Model predictions and experimental results by scenario. (A) In Dish Clean-Up, participants were told that all guests except for the red agent brought their plates (workload) to the sink(s) and then washed them. In Experiment 1, workload and group size were varied, while the number of bottlenecks was fixed at 1 sink. (B) In Salad Preparation, participants were told that all guests except for the red agent washed the carrots (workload) in the sink and then chopped them on the cutting board(s). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

A.1. Power analysis

To determine our target sample size, we conducted a simulation-based power analysis. This approach used parameter estimates drawn from pilot data for 40 subjects to simulate synthetic datasets across a range of sample sizes. Specifically, we simulated responses for our pre-registered linear mixed-effects model that included random intercepts for participants and fixed effects for our hypothesized predictors and their interactions. We then fit the same mixed-effects model to each simulated dataset and recorded the proportion of simulations (out of 100 per sample size) in which each coefficient reached statistical significance at or below 0.05. Once the sample size exceeded 300 subjects for Experiment 1, the power to detect our key fixed effects stabilized above 90% (higher than the commonly recommended 80% threshold). We repeated this analysis and found similar results above 90% for each experiment's pilot sample. We therefore chose $N = 300$ as the sample size for all of our experiments in this study.

A.2. Manipulating idle agent's effort by distance

In an earlier version of this experiment, participants were presented with a series of dinner party scenarios in which the idle agent's distance from the relevant workstation was randomized. Participants were asked to evaluate the acceptability of this agent's social loafing behavior. The preregistration can be found at https://aspredicted.org/1CP_G81. We found similar but weaker effects to Experiment 1 in regards to workload and group size. In the later iterations of these experiments found in this paper, we fixed distance to reduce variance.

We examined whether the idle agent's distance from the workstations affected subject judgments of idleness acceptability. An ANOVA on acceptability ratings across the three distance conditions (small, medium, large) revealed a significant main effect of distance, $F(2, 5044) = 12.57, p < 0.001$. As shown in Fig. A.7, participants deemed idleness more acceptable when the idle agent was a larger distance ($M = 5.77, SE = 0.06$) compared to medium ($M = 5.41, SE = 0.07$) or small ($M = 5.35, SE = 0.07$) distances. To account for random variation across participants, we also fit a linear mixed-effects model with random intercepts for participants and found that only the large

compared to medium distance reached significance (β = 0.36, SE = 0.07, 95% CI = [0.23, 0.49], $z = 5.29$, $p < 0.001$). The small compared to medium distance did not reach significance (β = −0.05, SE = 0.07, 95% CI = [−0.20, 0.09], $z = -0.72$, $p = -0.20$). These findings show a small but significant effect of the amount of effort required to complete a task on idleness acceptability (see Fig. A.7).

Data availability

The data and analysis code supporting the findings of this study are publicly available on GitHub at https://github.com/emieczkowski/IdleCollaborators and are also listed in the "Resources for Reproducibility" section.

References

Almaatouq, A., Alsobay, M., Yin, M., & Watts, D. J. (2021). Task complexity moderates group synergy. *Proceedings of the National Academy of Sciences*, *118*(36), Article e2101062118.

Almasi, G. S., & Gottlieb, A. (1994). *Highly parallel computing (2nd ed.)*. Benjamin-Cummings Publishing Co., Inc..

Amdahl, G. M. (1967). Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18-20, 1967, spring joint computer conference* (pp. 483–485).

Asch, S. E. (1940). Studies in the principles of judgments and attitudes: II. Determination of judgments by group and by ego standards. *The Journal of Social Psychology*, *12*(2), 433–465.

Baer, J.-L. (1973). A survey of some theoretical aspects of multiprocessing. *ACM Computing Surveys*, *5*(1), 31–80.

Baer, C., & Odic, D. (2022). Mini managers: Children strategically divide cognitive labor among collaborators, but with a self-serving bias. *Child Development*, *93*(2), 437–450.

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition*, *167*, 25–37.

Braun, T. D., Siegel, H. J., Beck, N., Bölöni, L. L., Maheswaran, M., Reuther, A. I., et al. (2001). A comparison of eleven static heuristics for mapping a class of independent tasks onto heterogeneous distributed computing systems. *Journal of Parallel and Distributed Computing*, *61*(6), 810–837.

Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., & Zelinsky, G. J. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, *106*(3), 1465–1477.

Brosnan, S. F. (2011). A hypothesis of the co-evolution of cooperation and responses to inequity. *Frontiers in Neuroscience*, *5*, 43.

Bruggen, A. (2015). An empirical investigation of the relationship between workload and performance. *Management Decision*, *53*(10), 2377–2389.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., et al. (2019). On the utility of learning about humans for human-ai coordination. *Advances in Neural Information Processing Systems*, *32*.

Cassidy, A. S., & Andreou, A. G. (2011). Beyond Amdahl's law: An objective function that links multiprocessor performance gains to delay and energy. *Institute of Electrical and Electronics Engineers. Transactions on Computers*, *61*(8), 1110–1126.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591–621.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. *vol. 24*, In *Advances in experimental social psychology* (pp. 201–234). Elsevier.

Cristian, F. (1991). Understanding fault-tolerant distributed systems. *Communications of the ACM*, *34*(2), 56–78.

Davis, R. I., & Burns, A. (2011). A survey of hard real-time scheduling for multiprocessor systems. *ACM Computing Surveys*, *43*(4), 1–44.

Dessein, W., & Santos, T. (2006). Adaptive organizations. *Journal of Political Economy*, *114*(5), 956–995.

Earley, P. C. (1989). Social loafing and collectivism: A comparison of the United States and the People's Republic of China. *Administrative Science Quarterly*, 565–581.

Flynn, E., Turner, C., & Giraldeau, L.-A. (2018). Follow (or don't follow) the crowd: Young children's conformity is influenced by norm domain and age. *Journal of Experimental Child Psychology*, *167*, 222–233.

Gerard, H. B., Wilhelmy, R. A., & Conolley, E. S. (1968). Conformity and group size. *Journal of Personality and Social Psychology*, *8*(1p1), 79.

Gerstenberg, T., Goodman, N., Lagnado, D., & Tenenbaum, J. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. *vol. 34*, In *Proceedings of the annual meeting of the cognitive science society*.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In *CogSci*.

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, *115*(1), 166–171.

Gerstenberg, T., Lagnado, D. A., et al. (2023). Making a positive difference: Criticality in groups. *Cognition*, *238*, Article 105499.

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? From expectations to responsibility judgments. *Cognition*, *177*, 122–141.

Griffiths, T. L. (2020). Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, *24*(11), 873–883.

Gustafson, J. L. (1988). Reevaluating Amdahl's law. *Communications of the ACM*, *31*(5), 532–533.

Hackman, J. R., & Morris, C. G. (1975). Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration. *Advances in Experimental Social Psychology*, *8*, 45–99.

Hawkins, R. D., Franke, M., Frank, M. C., Goldberg, A. E., Smith, K., Griffiths, T. L., et al. (2023). From partners to populations: A hierarchical Bayesian account of coordination and convention.. *Psychological Review*, *130*(4), 977.

Hennessy, J. L., & Patterson, D. A. (2011). *Computer architecture: a quantitative approach*. Elsevier.

Hill, M. D., & Marty, M. R. (2008). Amdahl's law in the multicore era. *Computer*, *41*(7), 33–38.

Ingham, A. G., Levinger, G., Graves, J., & Peckham, V. (1974). The Ringelmann effect: Studies of group size and group performance. *Journal of Experimental Social Psychology*, *10*(4), 371–384.

Jackson, J. M., & Harkins, S. G. (1985). Equity in effort: An explanation of the social loafing effect. *Journal of Personality and Social Psychology*, *49*(5), 1199–1206.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, *65*(4), 681–706.

Kerr, N. L., & Bruun, S. E. (1983). Dispensability of member effort and group motivation losses: Free-rider effects. *Journal of Personality and Social Psychology*, *44*(1), 78–94.

Kravitz, D. A., & Martin, B. (1986). Ringelmann rediscovered: The original article. *Journal of Personality and Social Psychology*, *65*, 681–706.

Kshemkalyani, A. D., & Singhal, M. (2011). *Distributed computing: principles, algorithms, and systems*. Cambridge University Press.

Kwon, J., Zhi-Xuan, T., Tenenbaum, J., & Levine, S. (2023). When it is not out of line to get out of line: The role of universalization and outcome-based reasoning in rule-breaking judgments. *PsyArXiv*.

Legare, C. H., & Wen, N. (2014). The effects of ritual on the development of social group cognition. *International Society for the Study of Behavioral Development*, *2*(66), 9–12.

Littlepage, G. E. (1991). Effects of group size and task characteristics on group performance: A test of Steiner's model. *Personality and Social Psychology Bulletin*, *17*(4), 449–456.

Mao, A., Mason, W., Suri, S., & Watts, D. J. (2016). An experimental study of team size and performance on a complex task. *PloS One*, *11*(4), Article e0153048.

Marjieh, R., Gokhale, A., Bullo, F., & Griffiths, L. T. (2024). Task allocation in teams as a multi-armed bandit. In *ACM collective intelligence*.

Mas, A., & Moretti, E. (2009). Peers at work. *American Economic Review*, *99*(1), 112–145.

McAuliffe, K., & Thornton, A. (2015). The psychology of cooperation in animals: an ecological approach. *Journal of Zoology*, *295*(1), 23–35.

Messeri, C., Bicchi, A., Zanchettin, A. M., & Rocco, P. (2022). A dynamic task allocation strategy to mitigate the human physical fatigue in collaborative robotics. *IEEE Robotics and Automation Letters*, *7*(2), 2178–2185.

Olson, M., Jr. (1971). *The logic of collective action: public goods and the theory of groups*: *vol. 124*, harvard University Press.

Petty, R. E., Harkins, S. G., Williams, K. D., & Latane, B. (1977). The effects of group size on cognitive effort and evaluation. *Personality and Social Psychology Bulletin*, *3*(4), 579–582.

Quillien, T., & Lucas, C. G. (2023). Counterfactuals and the logic of causal selection.. *Psychological Review*.

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2012). No third-party punishment in chimpanzees. *Proceedings of the National Academy of Sciences*, *109*(37), 14824–14829.

Ringelmann, M. (1913). Research on animate sources of power: The work of man. *vol. 12*, In *Annales de l'Institut National Agronomique* (pp. 1–40).

Roberts, S. O., Gelman, S. A., & Ho, A. K. (2017). So it is, so it shall be: Group regularities license children's prescriptive judgments. *Cognitive Science*, *41*, 576–600.

Schmidt, M. F., & Tomasello, M. (2012). Young children enforce social norms. *Current Directions in Psychological Science*, *21*(4), 232–236.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*, 487–510.

Simms, A., & Nichols, T. (2014). Social loafing: A review of the literature. *Journal of Management Policy and Practice*, *15*(1), 58–67.

Siswanto, S., Supriyanto, A., Ni'mah, U., Asnawi, N., & Wekke, I. (2019). Does a workload influence the performance of bank employees? *Management Science Letters*, *9*(5), 639–650.

Steiner, I. D. (1972). *Group process and productivity*. Academic Press.

Tabibnia, G., & Lieberman, M. D. (2007). Fairness and cooperation are rewarding: evidence from social cognitive neuroscience. *Annals of the New York Academy of Sciences*, *1118*(1), 90–101.

Turner, C. R., Nielsen, M., & Collier-Baker, E. (2014). Groups' actions trump injunctive reaction in an incidental observation by young children. *PloS One*, *9*(9), Article e107375.

Vélez, N., Christian, B., Hardy, M., Thompson, B. D., & Griffiths, T. L. (2023). How do humans overcome individual computational limitations by working together? *Cognitive Science*, *47*(1), Article e13232.

Williams, K., Harkins, S. G., & Latané, B. (1981). Identifiability as a deterrant to social loafing: Two cheering experiments. *Journal of Personality and Social Psychology*, *40*(2), 303–311.

Wu, S. A., & Gerstenberg, T. (2024). If not me, then who? Responsibility and replacement. *Cognition*, *242*, Article 105646.

Wu, S. A., Sridhar, S., & Gerstenberg, T. (2023). A computational model of responsibility judgments from counterfactual simulations and intention inferences. *vol. 45*, In *Proceedings of the annual meeting of the cognitive science society*.

Xiang, Y., Landy, J., Cushman, F. A., Vélez, N., & Gershman, S. J. (2023). Actual and counterfactual effort contribute to responsibility attributions in collaborative tasks. *Cognition*, *241*, Article 105609.

Xiang, Y., Vélez, N., & Gershman, S. J. (2023). Collaborative decision making is grounded in representations of other people's competence and effort. *Journal of Experimental Psychology: General*, *152*(6), 1565.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, Article e1.

Yu, D., & Thompson, B. (2024). People balance joint reward, fairness and complexity to develop social norms in a two-player game. *vol. 46*, In *Proceedings of the annual meeting of the cognitive science society*.