UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Do Large Language Models Resolve Fairness-Efficiency Trade-offs Like People?

Permalink

https://escholarship.org/uc/item/65m873sn

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Li, Cheryl Y
Mieczkowski, Elizabeth
Valera, Veronica
et al.

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Do Large Language Models Resolve Fairness–Efficiency Trade-offs Like People?

Cheryl Li (cl7672@princeton.edu)[1,*], **Elizabeth Mieczkowski**[2,*], **Veronica Valera**[2], **Natalia Vélez**[3] **& Thomas L. Griffiths**[2,3]

[1]Department of Electrical and Computer Engineering, Princeton University
[2]Department of Computer Science, Princeton University
[3]Department of Psychology, Princeton University
[*]Equal contribution

Abstract

With rapid improvements in reasoning and accuracy, large language models (LLMs) have gained more prominent roles as decision-makers in social and organizational settings. This makes it important to understand whether they adhere to the same values as people about division of labor. Here, we test whether LLMs and humans are aligned on task allocation problems, in which a set of tasks need to be assigned to collaborators. These problems present a trade-off between efficiency (optimizing for metrics like output and completion time) and fairness (where each collaborator must complete roughly an equal or equitable share of the overall workload). In two experiments we find that people and LLMs vary in how closely their stated preferences between efficiency and fairness align. Humans and LLMs frequently diverged from their stated preferences and settled on similar allocations when actively determining an allocation on their own. Our work identifies a value-action gap in both humans and LLMs that influences the degree to which LLMs align with social preferences.

Keywords: collaboration; fairness; efficiency; large language models; social preferences

Introduction

The ability to collaborate underlies the success of most innovations in history (Henrich, 2016; Vélez et al., 2023). However, collaboration itself requires solving many additional problems. One key challenge is task allocation: how to assign jobs to different collaborators given varying task demands, skills, and preferences. While there are many ways to distribute responsibilities depending on a group's priorities, task allocation commonly poses a tension between *efficiency*, or minimizing the total time required to complete all tasks, and *fairness*, which itself can reflect multiple principles such as equality and equity (Colquitt et al., 2001; Corbit et al., 2021).

The fairness-efficiency trade-off becomes especially salient when collaborators differ in skill or productivity (Liang & Wang, 2025). While allocating more work to a highly skilled collaborator improves total output, it leads to inequalities in workload. Failing to properly balance these trade-offs can undermine group cohesion, satisfaction, and performance (Rea et al., 2021). Understanding how people resolve these trade-offs is therefore crucial for understanding how people collaborate, and for for designing useful artificial collaborators.

Recently, large language models (LLMs) have been deployed in social and economic roles, including workforce scheduling, managerial decision support, and employee evaluation (Sapkota et al., 2026). As a result, LLMs are beginning to make and influence the same kinds of allocation decisions that human collaborators routinely face (Murugadoss et al., 2025). Although recent work suggests that LLMs can solve decision-making tasks similarly to humans (Binz & Schulz, 2023), other studies show systematic divergences between LLMs and humans in distributive preferences, such as monetary allocation (Hosseini & Khanna, 2025). It therefore remains unclear whether LLMs align with human social preferences in task-allocation decisions that trade off fairness and efficiency under realistic productivity constraints. As these models become embedded in systems that mediate interactions between people, understanding how LLMs reason about fairness–efficiency trade-offs is increasingly important.

In this paper, we examine how humans and LLMs allocate tasks when faced with explicit fairness-efficiency trade-offs. In Experiment 1, we present participants and a suite of LLMs with choices between pairs of task allocations across a range of collaborative vignettes. We find that human preferences vary with the context of the vignette, but overall tend to favor fair allocations over more efficient ones. In contrast, LLMs exhibit substantial heterogeneity in their alignment with human fairness judgments when presented with the same decisions. In Experiment 2, we examine how humans and LLMs construct task allocations when optimizing for fairness or efficiency. Although people and LLMs rely on different operational definitions of fairness, their fair allocations often converge due to small, permissible deviations from their preferred solution to partially satisfy alternative metrics. By studying how LLMs make decisions regarding these trade-offs, we aim to understand whether these judgments align with those of humans.

Background

Human collaboration and fairness

The benefits of teamwork depend jointly on the task itself (Mieczkowski et al., 2025), how fairly group members are treated (Chang et al., 2020; Liang & Wang, 2025), and how fairness is defined. Children, for example, develop fair resource allocations that take into consideration merit (Koskuba et al., 2018; Rizzo et al., 2016) or the welfare of others (Corbit et al., 2021) depending on the context. While people also perceive equity-based task allocations as fair when tasks vary in difficulty (Liang & Wang, 2025), they tend to produce allocations that prioritize the equal division of labor in a visual search task (Wahn & Wiese, 2025). In collaborations with a

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

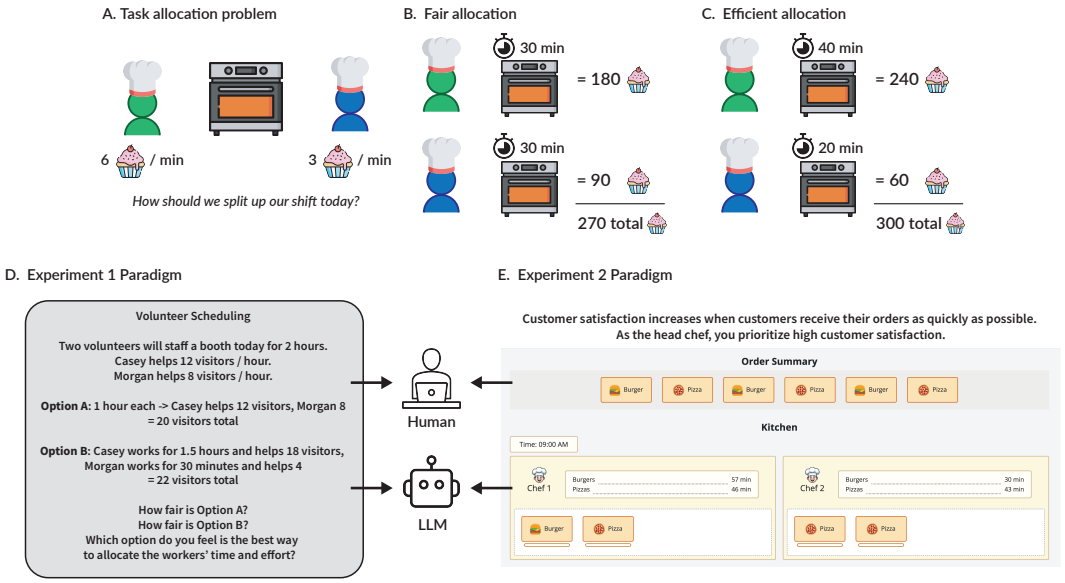

Figure 1: **Experimental framework. A.** In a task allocation problem, a manager needs to divide a task between collaborators. For example, how should two chefs split up their shift? **B.** A *fair* allocation might require both chefs to work the same amount of time, even though one is faster. **C.** An *efficient* allocation may instead exploit the faster worker, asking them to work more. **D.** In Experiment 1, participants and LLMs were presented with collaborative vignettes. In each scenario, they rated how fair two given allocations were, and which they preferred. **E.** In Experiment 2, participants and LLMs were asked to directly make these allocations themselves while optimizing for either fairness ("employee satisfaction") or efficiency ("customer satisfaction").

discrete set of tasks, metrics for equality can be further distinguished by differences in group members' skills, the amount of tasks, and specialization (Chang et al., 2020). However, in motor movement tasks, people systematically preferred maximizing group efficiency over fairness (Strachan & Török, 2020). Furthermore, fairness perceptions depend on the process through which an allocation is determined. People often perceive algorithmic decision-making, including the use of LLMs, as unfair because it reduces accountability (Binns et al., 2018) and limits worker' control over the process (Colquitt et al., 2001). These variations complicate our understanding of how people perceive fairness in different circumstances, especially when these decisions are delegated to LLMs.

Social preferences in LLMs

In addition to imitating human judgments across psychological experiments (Binz & Schulz, 2023; Dillion et al., 2023), LLMs frequently align with human preferences along many dimensions, including safety (Guan et al., 2025), moral values (Tennant et al., 2025), and social biases (Acerbi & Stubbersfield, 2023). Recent work, for example, shows that LLMs replicate human tendencies to emphasize socially desirable traits when under evaluation (Salecha et al., 2024). Value alignment is crucial for LLMs to collaborate closely with humans (Khamassi et al., 2024). However, the ability of LLMs to approximate these preferences in humans is brittle, and often depends on the prompt structure and context (Agarwal et al., 2024). Related work in resource allocation finds that LLMs tend to prioritize efficiency-based allocations contrary to human preferences for equitable solutions (Hosseini & Khanna, 2025). Furthermore, LLMs often demonstrate value-action gaps, in which a model's stated preference may fail to predict its action (Shen et al., 2025). Our paper extends these results to understand whether LLMs apply similar fairness preferences as humans to predictably create a fair allocation.

Balancing trade-offs in LLMs

A burgeoning area of research aims to understand how LLMs reason about trade-offs and optimize for multiple objectives (Liu et al., 2025; Q. Zhang et al., 2024). When presented with the option to produce honest or socially-acceptable responses LLMs were found to prioritize usefulness (Murthy et al., 2025). Faced with the classic dilemma between exploration and exploitation, LLMs reproduce a mix of directed and random exploration strategies often observed in humans (Z. Zhang et al., 2024). This ability to resolve conflicting values allows LLMs to reinforce an understanding of individual social values and their appropriate contexts (Liu et al., 2025).

Theoretical Framework

People often collaborate in order to improve collective efficiency, but doing so requires determining how responsibilities should be divided among group members. We formalize efficiency as the rate at which a group completes work:

$$\text{efficiency} = \frac{\text{group throughput}}{\text{completion time}} \quad (1)$$

Efficiency under sequential work

In some collaborative settings, workers contribute sequentially under a fixed time horizon. A common example is shift scheduling, in which N workers are allocated time intervals t_i within a total available time t_{total}, and complete work at rates r_i. In this scenario, efficiency is limited by:

$$\text{group throughput} = \sum_{i=1}^{N} r_i t_i \quad \text{s.t.} \quad \sum_{i=1}^{N} t_i \leq t_{\text{total}} \tag{2}$$

Efficiency under parallel work

In other settings, workers operate in parallel on a fixed set of tasks. Here, each worker i completes assigned tasks $A_i \subseteq T$, where task j requires time t_{ij}. Efficiency is determined by the time required to complete the entire workload:

$$\text{completion time} = \max_i \sum_{j \in A_i} t_{ij} \tag{3}$$

Fairness constraints on efficiency

In both regimes, purely efficiency-maximizing allocations may violate fairness considerations. We model fairness as a constraint on allocations. In sequential settings, fairness may be defined over time allocations:

$$F_{\text{time}} = \frac{t_1 - t_2}{t_1 + t_2}, \tag{4}$$

In parallel settings, fairness may instead be defined over task assignments:

$$F_{\text{equality}} = \frac{|A_1| - |A_2|}{|T|}, \tag{5}$$

or relative to worker capabilities:

$$F_{\text{capability}} = \frac{|A_1 \cap S_1|}{|S_1|} - \frac{|A_2 \cap S_2|}{|S_2|} \tag{6}$$

where S_i is the set of tasks at which worker i is fastest. For each metric, $F = 0$ indicates a "fair" allocation.

These formulations distinguish two classes of task allocation problems and formalize how efficiency and fairness trade off in each. Experiment 1 examines how humans and LLMs judge prespecified allocations in sequential work settings with a shared resource bottleneck, while Experiment 2 tests how agents construct allocations in parallel work settings.

Experiment 1: Comparing human and LLM fairness-efficiency preferences

When given the choice between a fair and efficient division of labor, do LLMs align with people? We first aim to understand human versus LLM preferences across a series of vignettes with a binary choice between a fair and efficient allocation.

Methods

Participants. We recruited 50 English-speaking adults from Prolific in exchange for monetary compensation ($1.60 for a 5-8 minute experiment). Four participants failed to complete attention checks, resulting in a final sample of 46 participants. The experiment was approved by the institutional review board. All participants provided informed consent.

Design and stimuli. We developed eight vignettes describing a task allocation problem with a shared resource ("sequential work"; Figure 1D). The vignettes span a range of collaborative contexts, such as running simulations on a computer, volunteering at a booth, or allocating work hours during a shift. In each scenario, the two workers differ in their task completion rates such that fairness is quantified by F_{time} and efficiency is determined by the workers' final output. Participants are presented with two alternative allocations: Option A schedules both workers for equal amounts of time (fair) and Option B assigns more time to the faster worker to increase total output (efficient). Each allocation specifies the resulting output. For each vignette, participants rated perceived fairness of each option on a 7-point Likert scale and indicated which they would prefer to implement. To capture relative fairness judgments, we also computed the difference between the ratings assigned to Options A and B (A–B), which quantifies participants' comparative assessment of the two allocations.

LLM Comparison. We selected six widely-used LLMs (non-reasoning: GPT-4o, Claude-Sonnet-4.5, Llama-3.1-70B; reasoning: GPT-5, Gemini-2.5-Pro, Grok-4-Fast). Each model was prompted with the same instructions and vignettes as human participants using a temperature of 0.5 and a seed of 1. We queried each model 30 times per vignette.

Results

People favor fair allocations unless they conflict with collective goals. Do people prefer task assignments that are efficient or fair? In this exploratory analysis, we first verified that our vignettes instantiated a genuine fairness–efficiency tradeoff (Figure 2A). Participants consistently rated the equally split allocation as more fair (mean difference = 3.38, $t(45) = 14.67$, $p < 0.01$). We then tested whether people preferred the fair or efficient allocation (Figure 2B). Participants favored the fair option in all but one vignette (mean proportion = 0.58, $t(45) = 1.9$, $p = 0.03$), in which participants were explicitly told that their team would fail to complete their goal if all collaborators worked equally.

LLMs exhibit varying alignment with human preferences. Do LLMs exhibit fairness preferences similar to those of people? We compared model predictions with human judgments along two complementary dimensions: (i) perceived fairness ratings (Table 1), and (ii) allocation preferences, quantified as the correlation between model and human choices across vignettes. Models differed substantially in how closely they matched human fairness judgments (Table 1). Llama-3.1-70B and Gemini-2.5-Pro showed the strongest alignment with people, followed by Claude-4.5. GPT-4o and Grok-4-Fast exhibited moderate alignment, while GPT-5 showed weaker and less reliable associations. Preference alignment showed a similar but noisier pattern: Gemini-2.5-Pro was the only model to exhibit a significant correlation with human choices

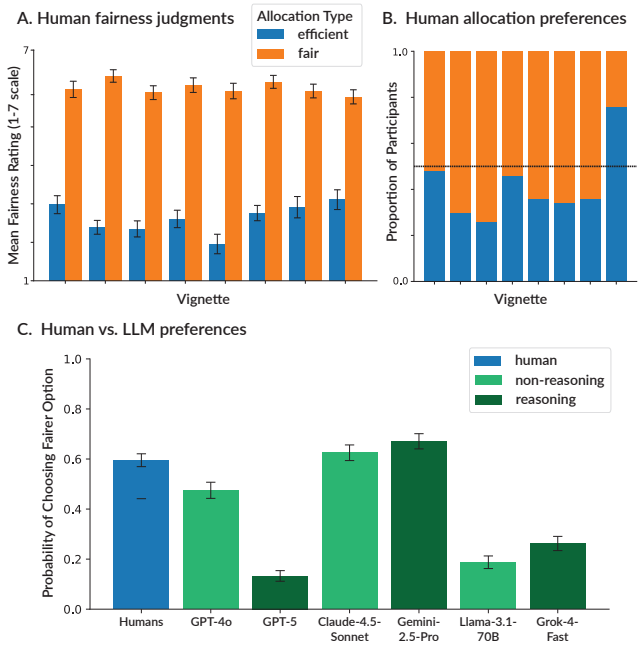

Figure 2: **Experiment 1 results. A.** Participants reliably rated equitable allocations as fairer than more efficient ones, confirming that the vignettes elicited a trade-off (error bars: SEM). **B.** In binary choices, participants tended to prefer the equitable allocation. **C.** In contrast, models exhibited heterogeneous preferences (error bars: SEM).

($r(6) = 0.80$, $p = 0.02$), whereas other models showed moderate or weak correspondence due to prioritizing efficiency ($r(6) = 0.21 - 0.56$, $p = 0.15 - 0.63$; Figure 2C). Notably, Llama-3.1-70B exhibited a strong dissociation between judgment and preference. Despite closely matching human fairness ratings, it demonstrated less alignment with preferences ($r(6) = 0.39$, $p = 0.34$). Follow-up analyses revealed a significant negative coupling between fairness and choice ($r(6) = -0.73$, $p = 0.04$), indicating that its decisions were systematically driven by efficiency rather than fairness.

Summary

These results indicate that people systematically prefer fairness, even when it sacrifices some efficiency that could be gained by exploiting differences in skill. LLMs largely differed in how well they captured these fairness judgments. While some models closely tracked human fairness ratings, they nevertheless prioritized efficiency more than people when selecting which allocation to implement.

Experiment 2: Comparing human and LLM task allocations

Do humans and LLMs make similar judgments when *constructing* an allocation that optimizes for fairness *or* efficiency? In Experiment 2, we test how humans and LLMs apply these fairness-efficiency preferences when dividing tasks between collaborators in a "parallel work" environment.

Table 1: Pearson correlations (r) between model-predicted fairness and human judgments, with p-values in parentheses.

Model	A	B	A–B
GPT-4o	0.81 (.02)	0.54 (.17)	0.76 (.03)
GPT-5	0.62 (.10)	0.37 (.37)	0.53 (.18)
Claude-4.5	0.91 (.00)	0.81 (.01)	0.91 (.00)
Gemini-2.5-Pro	0.96 (.00)	0.97 (.00)	0.97 (.00)
Llama-3.1-70B	0.99 (.00)	0.97 (.00)	1.00 (.00)
Grok-4-Fast	0.93 (.00)	.63 (.09)	0.87 (.00)

Methods

Participants We recruited 50 English-speaking adults from Prolific in exchange for monetary compensation ($5.00 for a 20-25 min experiment). Six participants failed to complete attention checks, resulting in a final sample of 44 participants. The experiment was approved by the Institutional Review Board, and all participants provided informed consent. The study was preregistered prior to data collection at https://aspredicted.org/yf9j8f.pdf.

Design and stimuli In the first two sections of the experiment, participants assigned a set of orders to a team of two chefs (Figure 1E). Participants were instructed to develop allocations that either minimized the chefs' total preparation time (efficiency) or balanced responsibilities across chefs (fairness). In the third section, participants were presented with a workload and two candidate allocations that balanced either the number of orders assigned to each chef (F_{equality}) or the number of orders related to a chef's strength ($F_{\text{capability}}$). Participants selected the allocation they judged to be fairest. Across all sections, we used a fixed set of ten workloads, each with ten orders. Workloads were constructed such that the optimally efficient allocation was not fair and yielded a total preparation time of 150 ± 10 time units.

LLM Comparison We prompted the same set of LLMs from Experiment 1 with a text-based version of the task. We report results for a temperature of 0.5 and a seed of 1, and again queried models 30 times over each workload and section.

Results

Fair allocations reduce efficiency. We first conduct a split half reliability analysis using an odd-even split by workload. Our results yield a strong correlation, indicating high reliability on reporting human performance under both efficient ($r_{sb} = 0.95$) and fair ($r_{sb} = 0.94$) conditions. Next, we fit a mixed linear model predicting completion time under both efficient and fair conditions (Figure 3A(i)). As expected, fair allocations typically led to longer completion times than efficient ones, a pattern observed across all agents and reaching

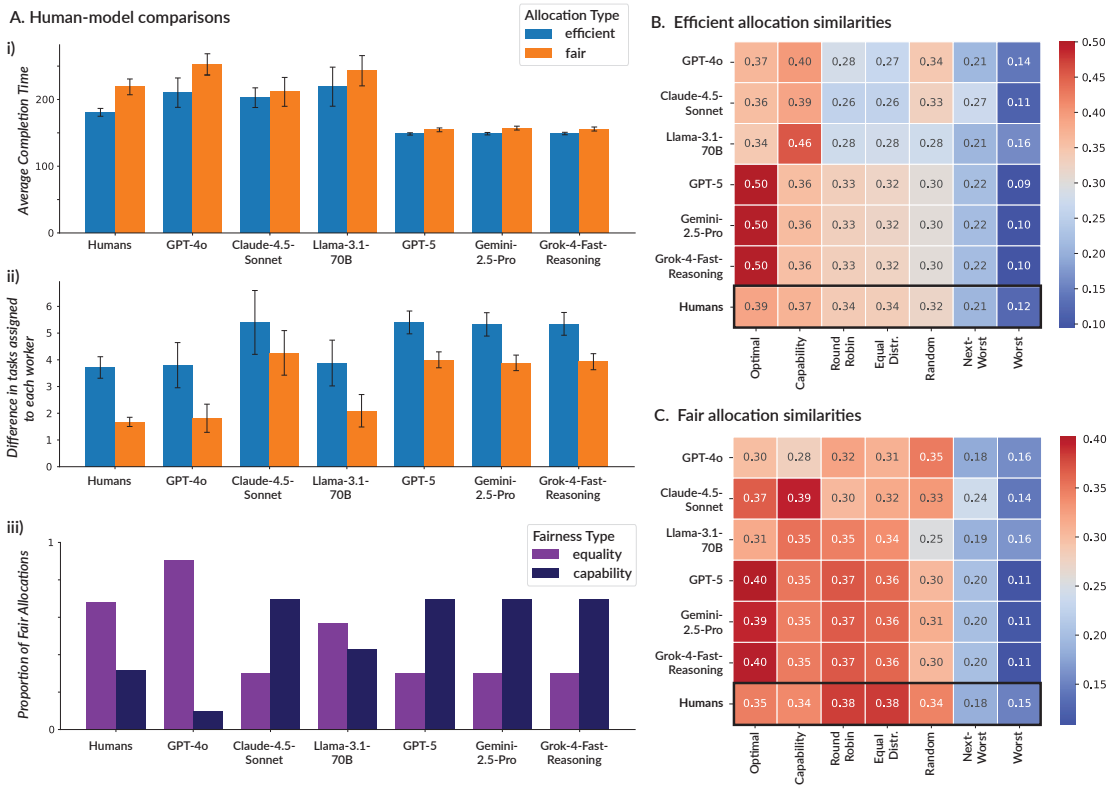

Figure 3: **Experiment 2 results. A.i.** When instructed to optimize for efficiency, both people and LLMs produced faster allocations than under fairness instructions. The efficiency–fairness gap was larger for people than for models (error bars indicate SEM). **A.ii.** Under fairness instructions, both people and LLMs generated allocations with smaller disparities across workers. In contrast, efficiency optimization led to disproportionate use of the faster worker. **A.iii.** When choosing between fair allocations, people tended to favor an equality-based definition, whereas models showed substantial variation. **B-C.** Heat maps show the similarity between human and model allocations relative to alternative algorithms.

significance for humans and most LLMs (Humans: $\beta = 38.15$, $z = 11.66$, $p < 0.01$; GPT-4o: $\beta = 42.27$, $z = 11.17$, $p < 0.01$; Llama-3.1-70B: $\beta = 23.89$, $z = 5.41$, $p < 0.01$; GPT-5: $\beta = 6.10$, $z = 12.92$, $p < 0.01$; Gemini-2.5-Pro: $\beta = 8.37$, $z = 6.80$, $p < 0.01$; Grok-4-Fast: $\beta = 6.76$, $z = 9.34$, $p < 0.01$); Claude-4.5 shows a non-significant effect in the same direction ($\beta = 8.70$, $z = 1.82$, $p = 0.07$). The increase in completion time confirms a non-trivial trade-off for problems that jointly optimize for fairness and efficiency.

Reasoning models find more efficient allocations than people. When instructed to optimize for efficiency, reasoning models (GPT-5, Gemini-2.5-Pro, and Grok-4-Fast) produced allocations that significantly outperformed those generated by people (GPT-5: mean difference $= -32.34$, $t(9) = -6.92$, $p < 0.01$; Gemini-2.5-Pro: -31.99, $t(9) = -6.86$, $p < 0.01$; Grok-4-Fast: -31.76, $t(9) = -6.60$, $p < 0.01$). Figure 3A(i) shows lower average completion times for reasoning models than for humans in both fair and efficient conditions.

Reasoning models favor efficiency even when prompted for fairness. Given their strong performance under efficiency, do reasoning models also find fairer allocations when prompted to do so? Interestingly, their "fair" allocations remained significantly more efficient than those produced by people (GPT-5: -26.24, $t(9) = -4.26$, $p < 0.01$; Gemini-2.5-Pro: -23.62, $t(9) = -4.31$, $p < 0.01$; Grok-4-Fast: -24.99, $t(9) = -3.80$, $p < 0.01$). Consequently, reasoning models generated allocations with larger disparities between workers even under fairness instructions (Figure 3A(ii)).

LLMs also differed in how they operationalized fairness when choosing between allocations that equalized task counts (F_{equality}) or balanced work according to chefs' strengths ($F_{\text{capability}}$). Both humans and GPT-4o consistently preferred allocations that equalized the number of assigned tasks (Humans: mean difference $= -0.37$, $t(9) = -4.26$, $p < 0.01$; GPT-4o: -0.80, $t(9) = -8.99$, $p < 0.01$). In contrast, reasoning models and Claude-4.5 deterministically favored a combination of equality-based and capability-based allocations for different workloads. These allocations consistently corresponded to the most efficient solution that equalized assigned task times across workers (F_{time}), and indicate that reasoning

models continue to place substantial weight on efficiency even when instructed to prioritize fairness. These results align with observations from the sequential work setting that LLMs prefer efficient allocations in scenarios where people prioritize fairness. This suggests that social preferences in both LLMs and humans transfer across regimes when balancing trade-offs between fairness and efficiency.

Humans and LLMs balance fairness-efficiency trade-offs. Neither people nor LLMs disregarded efficiency entirely when constructing fair allocations. Humans and GPT-4o increased equality in task assignments under fairness instructions (Humans: mean difference = 2.04, $t(9) = 7.63$, $p < .01$; GPT-4o: 1.99, $t(9) = 3.79$, $p < .01$), but rarely assigned identical numbers of tasks to each worker. As a result, human fair allocations remained substantially more efficient than allocations that enforce strict equality via random assignment (mean difference – 43.62, $t(9) = -4.84$, $p < .01$). Similarly, although their fairness judgments emphasized balancing time spent on tasks (F_{time}), reasoning models accepted modest increases in completion time relative to their own efficient solutions while simultaneously reducing disparities in task counts (GPT-5: mean difference = 1.40, $t(9) = 4.58$, $p < .01$; Gemini-2.5-Pro: 1.44, $t(9) = 4.88$, $p < .01$; Grok-4-Fast: 1.42, $t(9) = 4.58$, $p < .01$). Thus, both humans and LLMs treated fairness as a flexible constraint, adjusting allocations toward equality without fully sacrificing efficiency.

Reasoning models and humans produce similar fair allocations. Despite relying on different fairness criteria, humans and reasoning models often converged on similar fair allocations. When fairness was evaluated using task equality (F_{equality}), only reasoning models showed significant correlations with human allocations (Gemini-2.5-Pro: $r(8) = 0.73$, $p = 0.02$; GPT-5: $r(8) = 0.67$, $p = 0.03$; Grok-4-Fast: $r(8) = 0.67$, $p = 0.03$). Consistent with this convergence, fair allocations produced by reasoning models were significantly more similar to human fair allocations than those produced by non-reasoning models (mean difference = 0.07, $t(29) = 4.53$, $p < 0.01$). However, despite these similarities in task assignments, we note that differences in completion time and number of tasks assigned to each worker remain statistically between human and LLM fair allocations remain statistically significant. Overall, this suggests that different fairness preferences can lead to overlapping solutions once efficiency constraints are taken into account.

Summary

People and LLMs rely on different notions of fairness when allocating work; people prefer equalizing task counts, while LLMs, especially reasoning models, emphasize efficiency. Despite these differences, neither group treated fairness as a hard constraint; both people and reasoning models abandoned strict equality when efficiency gains were large. Together, these results show that both humans and LLMs negotiate fairness and efficiency rather than enforcing either in isolation.

Discussion

Collaboration requires decision-makers to balance conflicting goals and values. As LLMs become more involved in shaping these decisions, it is crucial to better understand both human preferences and how closely models reflect these tradeoffs. In this work, we examined how humans and LLMs navigate fairness–efficiency trade-offs across two task allocation paradigms. We find that people generally prioritize fairness over efficiency, as long as the collective objective is achieved. In contrast, LLMs varied substantially in their alignment with human judgments and frequently prioritized efficiency even when explicitly instructed to favor fairness. Despite these divergences, both humans and LLMs exhibited flexible tradeoff behavior when independently constructing their own task allocations. Strikingly, reasoning models produced fair allocations that not only resulted in greater collective efficiency than humans' efficient allocations, but also exhibited significantly greater similarity to humans' fair allocations.

Our results suggest that both humans and LLMs construct nuanced solutions that balance competing collaborative goals. They also highlight an important implication for LLM interpretability: prompting models to simply state their preferences is insufficient for predicting behavior in real-world settings where trade-offs arise. Instead, evaluating both behavior and how stated values translate into concrete decisions is essential for identifying alignment versus divergence. Our work clarifies this alignment in simplified task allocation problems and contributes to growing efforts to finetune LLMs to better reflect complex value trade-offs.

Our study has several limitations. In particular, our method cannot distinguish whether humans exhibit an intrinsic *preference* for fairness, or if they are naturally worse at efficient allocations. While recent work suggests that people can optimize for efficiency (Mieczkowski, Li, et al., 2026), future work is needed to clarify this gap. Additionally, we did not examine LLM reasoning traces, which should be characterized to better understand the internal representations underlying this tradeoff. More broadly, real-world collaborations are shaped by interpersonal relationships, individual objectives, and external pressures. Understanding how humans navigate these complexities, and developing LLMs that can jointly optimize competing objectives, will be essential for fostering collaborations that are both productive and fair.

Code Availability The stimuli, analysis code, and data used in this study are publicly available at https://github.com/cl7672/LLM_Human_FairEff_in_TaskAllocation.git.

Acknowledgments

This work was supported by the National Defense Science and Engineering Graduate (NDSEG) Fellowship Program to EM, and and ONR MURI N00014-24-1-2748 to TG.

References

Acerbi, A., & Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, *120*(44), e2313790120.

Agarwal, U., Tanmay, K., Khandelwal, A., & Choudhury, M. (2024). Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 6330–6340.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'it's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Chang, M. L., Pope, Z., Short, E. S., & Thomaz, A. L. (2020). Defining fairness in human-robot teams. *29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 1251–1258.

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O. L. H., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *The Journal of Applied Psychology*, *86*(3), 425–445.

Corbit, J., Brunschwig, V., & Callaghan, T. (2021). The influence of collaboration on children's sharing in rural India. *Journal of Experimental Child Psychology*, *211*, 105225.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600.

Guan, M. Y., Joglekar, M., Wallace, E., Jain, S., Barak, B., Helyar, A., Dias, R., Vallone, A., Ren, H., Wei, J., & et al. (2025). Deliberative alignment: Reasoning enables safer language models. *SuperIntelligence - Robotics - Safety and Alignment*, *2*(3).

Henrich, J. (2016). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Hosseini, H., & Khanna, S. (2025). Distributive fairness in large language models: Evaluating alignment with human values. *arXiv preprint arXiv:2502.00313*.

Khamassi, M., Nahon, M., & Chatila, R. (2024). Strong and weak alignment of large language models with human values. *Scientific Reports*, *14*(1).

Koskuba, K., Gerstenberg, T., Gordon, H., Lagnado, D., & Schlottmann, A. (2018). What's fair? How children assign reward to members of teams with differing causal structures. *Cognition*, *177*, 234–248.

Liang, J., & Wang, H. (2025). Is It Fair Enough? Supporting Equitable Group Work Assignment with Work Division Dashboard. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2480–2490.

Liu, A., Ghate, K., Diab, M., Fried, D., Kasirzadeh, A., & Kleiman-Weiner, M. (2025). Generative Value Conflicts Reveal LLM Priorities. *arXiv preprint arXiv:2509.25369*.

Mieczkowski, E., Li, C., Vélez, N., & Griffiths, T. L. (2026). People intuitively schedule tasks to improve collective efficiency. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *48*.

Mieczkowski, E., Turner, C., Vélez, N., & Griffiths, T. L. (2025). People evaluate idle collaborators based on their impact on task efficiency. *Cognition*, *264*, 106200.

Murthy, S. K., Zhao, R., Hu, J., Kakade, S., Wulfmeier, M., Qian, P., & Ullman, T. (2025). Using cognitive models to reveal value trade-offs in language models. *Submitted to The Fourteenth International Conference on Learning Representations*.

Murugadoss, B., Poelitz, C., Drosos, I., Le, V., McKenna, N., Negreanu, C. S., Parnin, C., & Sarkar, A. (2025). Evaluating the Evaluator: Measuring LLMs' Adherence to Task Evaluation Instructions. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(18), 19589–19597.

Rea, D., Froehle, C., Masterson, S., Stettler, B., Fermann, G., & Pancioli, A. (2021). Unequal but fair: Incorporating distributive justice in operational allocation models. *Production and Operations Management*, *30*(7), 2304–2320.

Rizzo, M. T., Elenbaas, L., Cooley, S., & Killen, M. (2016). Children's recognition of fairness and others' welfare in a resource allocation task: Age related changes. *Developmental Psychology*, *52*(8), 1307–17.

Salecha, A., Ireland, M. E., Subrahmanya, S., Sedoc, J., Ungar, L. H., & Eichstaedt, J. C. (2024). Large language models display human-like social desirability biases in big five personality surveys. *PNAS Nexus*, *3*(12), pgae533.

Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2026). AI Agents vs. Agentic AI: A Conceptual taxonomy, applications and challenges. *Information Fusion*, *126*, 103599.

Shen, H., Clark, N., & Mitra, T. (2025). Mind the Value-Action Gap: Do LLMs Act in Alignment with Their Values? *arXiv preprint arXiv:2501.15463*.

Strachan, J. W., & Török, G. (2020). Efficiency is prioritised over fairness when distributing joint actions. *Acta Psychologica*, *210*, 103158.

Tennant, E., Hailes, S., & Musolesi, M. (2025). Moral alignment for LLM agents. *The Thirteenth International Conference on Learning Representations*.

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), e13232.

Wahn, B., & Wiese, E. (2025). Play fair: Humans prefer an equal division of labor in a joint multiple object tracking task. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Zhang, Q., Duan, Q., Yuan, B., Shi, Y., & Liu, J. (2024). Exploring accuracy-fairness trade-off in large language models. *arXiv preprint arXiv:2411.14500*.

Zhang, Z., Wang, D., Chen, N., Mansur, R., & Sarhangian, V. (2024). Comparing Exploration–Exploitation Strategies of LLMs and Humans: Insights from Standard Multi-armed Bandit Experiments. *arXiv preprint arXiv:2411.14500*.