Adaptive use of vagueness to coordinate joint action

Dhara Yu, Bill D. Thompson
University of California, Berkeley
{dharakyu, wdt}@berkeley.edu

Abstract

"Let's share the load" is a much less helpful way to coordinate cleaning the house than *"You take the bedrooms, I'll do the kitchen"* - unless there are 12 bedrooms. Vague plans can be useful tools to support joint action, but using vagueness effectively is a difficult computational problem. Participants in a joint planning study selected between specific and vague plans to coordinate action across a systematic range of problems. In Experiment 1, participants deployed vague plans selectively, recognizing situations where the certainty of a bad plan is outweighed by the flexibility of a vague plan according to a probabilistic model of joint reasoning. In Experiment 2, participants with greater exposure to such situations endorsed vague plans when providing generic testimony to future actors. Our results highlight an understudied but potentially powerful dimension of human joint planning: strategic use of vague construals.

Keywords: vagueness; joint planning; testimony

Introduction

A hallmark of human intelligence is the ability to coordinate joint action through natural language (Tomasello, Carpenter, Call, Behne, & Moll, 2005). Joint plans expressed in language share similarities with computer programs in that both can convey algorithmic, procedural sequences of actions (D'Andrade, 1981). However, natural language plans differ from computer programs in a key way: plans expressed in language are often ambiguous, yet they can nonetheless be "interpreted" by another person (Acquaviva et al., 2022).

This distinction highlights the range of strategies people can employ to coordinate joint action. On one end, they can issue exact procedural operations that more closely resemble computer programs. Imagine a team of archaeologists excavating a site; the director prescribes a precise series of actions (*"use a trowel to dig for the first 10 cm and then switch to a brush and a handpick"*) for fixed zones. This kind of plan supports reliable coordination in specific contexts, but is less useful in other situations (e.g. the archeologists change sites).

On the other end, plans can convey broad goals or general constraints for *determining* useful joint actions. Instead of exact procedural instructions, the director could tell team members to choose tools that avoid damage to artifacts, and focus their search on areas near standing structures. This type of communication is possible because people can use social reasoning to inform their actions (Ho, Saxe, & Cushman, 2022), and enrich the literal meaning of language (Grice, 1975; Piantadosi, Tily, & Gibson, 2012; Goodman & Frank, 2016), including vague language for which the meaning of a word itself is context-dependent (Lassiter & Goodman, 2017). Plans of this nature offer a powerful mechanism for conveying abstract knowledge (Tessler & Goodman, 2019) and coordinating action in more dynamic situations, but at greater risk of misunderstanding and coordination failure.

Choosing how specific to be when planning with other people poses a difficult computational problem. The problem arises in real-time interactions when communicating with a partner in a specific scenario and in the broader setting of transmitting accumulated knowledge to other people. Prior work has examined how people solve some of the core tasks implicated in this problem. People can decompose a joint task into individual actions (Wu et al., 2021; Gordon, Knoblich, & Pezzulo, 2023; Török, Pomiechowska, Csibra, & Sebanz, 2019), infer another agent's intentions in context from underspecified instructions (Stacy et al., 2021; Zhi-Xuan, Ying, Mansinghka, & Tenenbaum, 2024), consider other agents' perspectives (Hawkins, Gweon, & Goodman, 2021), and craft natural language instructions to convey abstract conceptual information (McCarthy, Hawkins, Wang, Holdaway, & Fan, 2021; Sumers, Ho, Hawkins, & Griffiths, 2023).

However, it remains unclear how effectively people arbitrate between specific and vague construals in joint-planning problems, and whether human reasoning aligns with the principles we have described. Understanding this ability could illuminate fundamental principles of adaptive collaboration. However, key questions have been difficult to address because existing multi-agent tasks do not systematically expose situations where vagueness could be useful. More generally, the value of vagueness in this setting is difficult to quantify.

Here we take steps towards assessing how effectively people leverage vagueness in joint planning by developing an experimental paradigm to study a simplified version of the problem. In Experiment 1, we examined how people interpret different natural language plans in a grid navigation task, and examined their choices of plans to use across a systematically varying range of scenarios. People preferentially chose vague plans in contexts where the interpretive flexibility of a vague plan was more valuable than the certainty of a specific but suboptimal plan. Plan selections were consistent with a model that instantiates probabilistic reasoning over another agent's interpretation of a plan. In Experiment 2, we investigated how people integrate knowledge from a distribution

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

Figure 1: Order of assignment can affect joint cost. Left: trajectories consistent with the blue-first assignment (the blue agent picks the closest item to it and the red agent takes the remaining item). Right: the red-first assignment.

of decision settings into generic testimony for future participants; we found that participants who experienced more problems where the flexibility of the vague plan was advantageous were systematically more likely to endorse the value of a vague plan during testimony. These results illustrate potential mechanisms through which abstract strategies for joint action may emerge and propagate through social transmission.

Experiment 1

Task overview

Participants were informed they would participate in a virtual scavenger hunt with a partner, in a simple grid environment. Each player's location is represented with an avatar; the participant is the blue agent; their partner is the red agent. Participants were told they need to collect a total of two specified checklist items located in the grid, for a range of grids. They were also informed that their bonus payments for participation are tied to the number of steps they individually take within the grid, incentivizing efficient trajectories. Participants selected a plan for each grid. They were told that this plan would be sent to their partner.

All sorts of plans could be useful in this task, from concrete object-level plans (*"You take the coin"*) or spatial plans (*"You cover the left half"*), to more abstract relational plans (*"I'll take the item that's closest to me"*) or general goals (*"Let's pick the fairest assignment"*). We chose to focus here on two relational plans, setting up a simple distinction in vagueness while preserving comparable algorithmic structure. Specifically, participants chose between a **vague** plan and a **specified** plan. The vague plan stated: *"We each take one item, picking the item that's reasonably close"*. The specified plan stated: *"We each take one item; I (blue) will pick the item closest to me first and then you (red) take the remaining item"*.

Both plans prescribe a 1-to-1 assignment between a player and an item (and exclude assignments of one player to both items). The specified plan provides an exact procedure to determine the assignment (compute the distances between the 2 items and the blue agent, and retrieve the item with the shorter associated distance if the blue agent, or the other item if the red agent). In contrast, the vague plan does not prescribe an explicit rule for determining who picks what. The vague plan could be consistent with the 1-to-1 assignment in which the blue agent selects the closer item to it and the red agent takes the remaining item (which we will refer to as the blue-first assignment). This assignment is entailed by the **specified** plan. It could *also* be consistent with the 1-to-1 assignment in which the partner red agent first selects its closer item, and the participant takes the remaining item (the red-first assignment). This assignment is entailed by a variant of the specified plan, that prescribes that the red player chooses first; we will distinguish this as the **specified** plan with the red-first order. Critically, the order in which participants select the closer item can substantially affect the amount of steps that must be taken to retrieve all the items in the grid (Figure 1).

Grid structure Each participant selected a plan for each of 12 grids. There were 6 categories of grids (2 grids per category). These grids were chosen to reveal the contingencies that affect preferences for the vague or specified plan. The grids vary on the axes of how much each player pays under a particular assignment (individual cost), the gap in cost paid between the two players (fairness), and the combined cost paid by both players (joint cost). Inducing conflicts between these features creates situations in which the assignment prescribed by the vague plan is ambiguous, as participants have to reason about what is entailed by "reasonably close".

***No uncertainty* grids**: In these grids, each of the two items was closer in proximity to a unique player. This means the blue-first assignment and the red-first assignment are the same, and consequently, the specified plan and the vague plan should entail the same assignment.

***Blue-first better* grids:** Here the blue-first assignment is strictly better for the participant (the blue player), as it results in a lower cost incurred, a more fair outcome and a lower joint cost, compared to the red-first assignment. We predicted that participants would prefer the specified plan (blue-first).

***Red-first better* grids:** Here the red-first assignment results in a fairer outcome with lower joint cost, at the expense of only a few additional steps for the participant (compared to the blue-first assignment). We predicted that participants would prefer the vague plan, which can be construed by the partner as consistent with the red-first assignment.

***Blue-first more fair + higher joint cost* grids:** These grids induce a direct conflict between fairness and joint cost. The blue-first assignment results in a smaller cost gap between participants but a higher joint cost. Equivalently, the red-first assignment results in a larger cost gap but a lower joint cost.

***Blue-first less fair + lower combined cost* grids:** These grids also induce a conflict between fairness and joint cost. However, following the blue-first assignment yields a larger cost gap and a lower joint cost (equivalently, the red-first assignment yields a smaller cost gap and a higher joint cost).

***Coordination challenge* grids:** These grids introduce ambiguity through a different mechanism. Each of the two players is an equal distance away from both items, so both possible 1-to-1 assignments of players to items yield the same individual costs, cost gaps, and joint costs: there is no cost-based reason to prefer one assignment over the other.

Computational framework

Our model captures the hypothesis that the value of a plan (in a specific grid) is given by the utility of following that plan, minus the cost of coordination failure, and that the calculation of utility varies across plan types due to differences in their interpreted meaning. Formally, $P = \{P_{\text{blue}}, P_{\text{red}}\}$ is the set of players and $I = \{C, T\}$ is the set of items (coin, trophy). Let $\mathcal{R} = \{\rho_1, \rho_2, \rho_3, \rho_4\}$, where ρ_i is a partition of items to players (an *assignment*). g is the grid configuration and σ represents a natural language plan (from P_{blue} to P_{red}). $P(\rho|\sigma, g)$ is the probability of pursuing assignment ρ under plan σ in grid g, and $C(\rho, g)$ is the joint cost of both players' actions under assignment ρ in grid g. The perceived value of plan σ in grid g is:

$$V(\sigma, g) = -w_S \cdot \left[\sum_{\rho \in \mathcal{R}} P(\rho|\sigma, g) \cdot C(\rho, g) \right] - w_F \cdot F(g).$$

This framework presumes rational action, in that it relates the value of a plan for a grid to the number of steps taken. w_S controls how strongly the decision-maker weighs the expected costs under successful outcomes (both players converge on the same assignment) when evaluating a plan; w_F represents the weight assigned to the expected costs for a coordination failure (both players initially go for the same item and then one has to divert to retrieve the remaining object). To capture the intuition that the cost of failure is related to the joint distance that would have to be traversed for successful outcomes, we set the expected cost of coordination failure $F(g) = 0.8 \cdot C(\rho^+, g)$ where ρ^+ is the costliest assignment.

The specified plan σ_S implies $P(\rho^*|\sigma_S, g) = 1, P(\rho \neq \rho^*|\sigma_S, g) = 0$, where ρ^* satisfies:

$\rho^* = \{P_{\text{blue}} : \{C\}, P_{\text{red}} : \{T\}\}$ if C is closer to P_{blue}, else swap

The probability of selecting the assignment in which the blue player selects its closest item is 1; the probability of all other assignments is 0. For clarity we will denote $\rho^* = \rho_{\text{blue-first}}$. We assume σ_s rules out coordination failure ($w_F = 0$), so $V(\sigma_S, g) = -w_S \cdot C(\rho_{\text{blue-first}}, g)$.

The vague plan σ_V implies multiple assignments may have nonzero probability, i.e. $P(\rho|\sigma_V, g) > 0$ for multiple ρ. Calculation of $V(\sigma, g)$ requires a way to compute these probabilities. Our goal here is not to identify a single model for these computations. $P(\rho|\sigma, g)$ can be viewed generically as a meaning function which maps utterances (a natural language plan) to world states (assignment of players to items) that could in principle be computed using a model of pragmatic inference of vague adjectives (i.e. inferring a truth-conditional threshold value for "reasonably close"; Lassiter and Goodman (2017)) or other probabilistic pragmatic models (Goodman & Frank, 2016). Here we implement a simple cost-based *joint-agent* (Kleiman-Weiner, Ho, Austerweil, Littman, & Tenenbaum, 2016) heuristic that approximates the outcome of interactive social reasoning. We assume that the vague plan rules out the two assignments where one player takes both items, and that the strength of one's belief that the vague plan would bring about one of the remaining assignments is *inversely proportional* to its joint cost, leading to:

$$P(\rho_{\text{blue-first}}|\sigma_v, g) = \frac{1}{1 + \exp(-k \cdot \Delta C)}$$

$$\Delta C = C(\rho_{\text{blue-first}}, g) - C(\rho_{\text{red-first}}, g)$$

and $P(\rho_{\text{red-first}}|\sigma_v, g) = 1 - P(\rho_{\text{blue-first}}|\sigma_v, g)$.

Finally, the probability of selecting plan σ is given by a softmax decision rule: $P(\sigma, g) \propto \exp(V(\sigma, g))$.

Experimental procedure

Task familiarization Participants viewed instructions explaining the calculation of cost in the game, and completed a short comprehension quiz to proceed before completing a series of task familiarization trials. These trials acquainted participants with scenarios in which different joint plans prescribe different assignments of players to items. Participants were shown one grid configuration and a natural language plan. Participants selected a set of actions that are consistent with the plan. Actions were visualized as player trajectories to items; since there are 2 players and 2 items, there are 4 possible assignments (since 1 player could collect both items). Participants could select more than one assignment per plan.

Each player selected an action set for 3 joint plans (the vague plan, the specified plan, and the specified plan with the red-first order) across 4 different grids (12 trials total). The 4 grids included 2 grids where the order in which agents select their closer item does not change the prescribed assignments of players to objects (no uncertainty grid type), and 2 grids where the order did affect the assignment (one where blue selecting first would lead to a lower joint cost, and one where red selecting first would lead to a lower joint cost).

Main trials After task familiarization, participants observed a series of scavenger hunt grids and for each grid, picked a plan to send to their partner. The grids were the same 12 described earlier, presented in a randomized order. After completing these trials, participants were informed they had not been paired with a partner and the experiment ended.

Participants We pre-registered[1] a target sample size of 50 participants, recruited from Prolific. We excluded from analysis participants who failed the quiz or did not complete all trials, leaving a total of $N = 40$. Participants were paid a base rate of $2.50 and were incentivized with a performance-based bonus to select helpful plans. Since they were not actually paired with another player, we paid all participants an additional $1.00. The task took a median time of 16 minutes (a median hourly rate of $13/hr).

Behavioral results

Systematic sensitivity to grid structure (and value) in plan interpretation Participants' interpretations of natural language plans were systematically related to 1) the structure of the grids, and 2) the implied costs. To evaluate participants'

[1]Link to pre-registration: https://osf.io/jx3bh/?view_only=0a8b5eb5983a491c848bd3c36450ab67

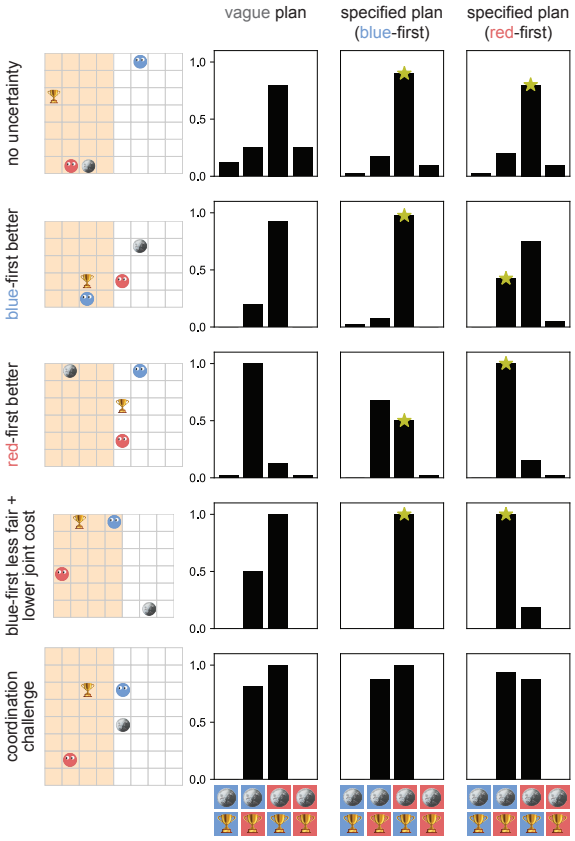

Figure 2: Participants' interpretations of different plan types across a variety of grids. Each bar represents the frequency that participants marked an assignment of players to items as consistent with the given plan. The star indicates the single true assignment under a literal interpretation of the plan.

understanding of the plans, we analyzed responses to the task familiarization trials, since participants were asked to parse plans to all the possible player-item assignments consistent with the plan. Task familiarization trials yielded data for 3 of the 6 grid types. We ran an additional study to collect data for the 3 remaining grid types in which additional participants ($N = 16$) performed the same task for the 3 plan types.

Figure 2 shows participants' interpretations.[2] In grids where one of the possible 1-to-1 assignments was clearly better (the no uncertainty, blue-first better and the red-first better grids, rows 1 - 3), participants overwhelmingly interpreted the vague plan (column 1) as consistent with the better assignment. On the other hand, for grids in which the preferred 1-to-1 assignment was more ambiguous, either because of a tension between fairness and joint cost (blue-first more fair + higher joint cost, row 4), or because both 1-to-1 assignments resulted in equally fair and costly outcomes (coordination challenge, row 5), participants indicated that multiple assignments were consistent with the vague plan. Participants were generally accurate in identifying the single assignment implied by the specific plans (blue-first or red-first, columns 2 & 3); however, their responses were more variable in cases where the prescribed ordering resulted in a worse outcome compared to the opposite ordering (the red-first order for the blue-first better grid: column 3, row 2; and the blue-first order for the red-first better grid: column 2, row 3). These results confirmed that participants were systematically sensitive to the differences between plan types and the task and grid contingencies, but also showed that participant interpretations were influenced by value computations.

Selective use of the vague plan Plan selections were systematically related to the value of specificity vs. vagueness in context. Participants overwhelmingly selected the specified plan for the blue-first better grids, consistent with the idea that they recognized that the assignment prescribed by specified plan (which is the blue-first assignment) is better than the alternative assignment in that context. A Bayesian logistic regression revealed they were significantly more likely to select the vague plan for the no uncertainty grids ($\beta = 2.19$, 95% credible interval (CrI) $= [1.45, 2.98]$); the red-first better grids ($\beta = 2.65$, CrI$= [1.94, 3.43]$); the blue-first more fair + higher joint cost grids ($\beta = 2.30$, CrI$= [1.56, 3.10]$); and the coordination challenge grids ($\beta = 1.26$, CrI$= [0.53, 2.03]$). The high selection rate of the vague plan in the red-first better grids suggests that people recognized that the assignment under the specified plan was suboptimal, and that the vague plan could be reliably parsed to prescribe the red-first assignment. The plan interpretation data (Figure 2) for the blue-first less fair + lower joint cost grid showed that participants recognized that multiple assignments are consistent with the vague plan; one explanation for the high rate of specified plans could be that people knew that the vague plan was ambiguous, and even if the assignment under the blue-first plan was imperfect for reasons of fairness, it solved the coordination problem.

Model results

We set $k = 4$ and $w_F = 1$ leaving one free parameter: w_C.[3] To make these parameters easier to interpret, we rescaled each $C(\rho, g)$ to a value between 0 and 1 by dividing by $\max(C(\rho_{\text{blue-first}}, g), C(\rho_{\text{red-first}}, g))$. To fit w_C we searched over the range $w_C \in ([0, 0.2, ..., 1] \cup [1, 2, ..., 20])$. We evaluated model fit to the data by computing the Jensen-Shannon divergence between the true and predicted plan distributions for each grid, and then taking the mean across all grids.

Figure 3 shows alignment (mean JSD = 0.058) between the predictions of the best-fitting model ($w_C = 14$) and the experimental data. The large inferred value of w_C relative to w_F suggests that people were highly sensitive to the expected costs of plans, relative to the risk of coordination failure.

Although the model captured the key patterns, there were a few areas of divergence. The model predicted that partici-

[2] Results from the blue-first more fair + lower joint cost grids are omitted for space, but demonstrate a similar pattern to blue-first less fair + higher joint cost grids.

[3] In cases where the preferred 1-to-1 assignment is assumed to be unambiguous, i.e. in the no uncertainty case, w_F is set to 0 because there is no risk of coordination failure.

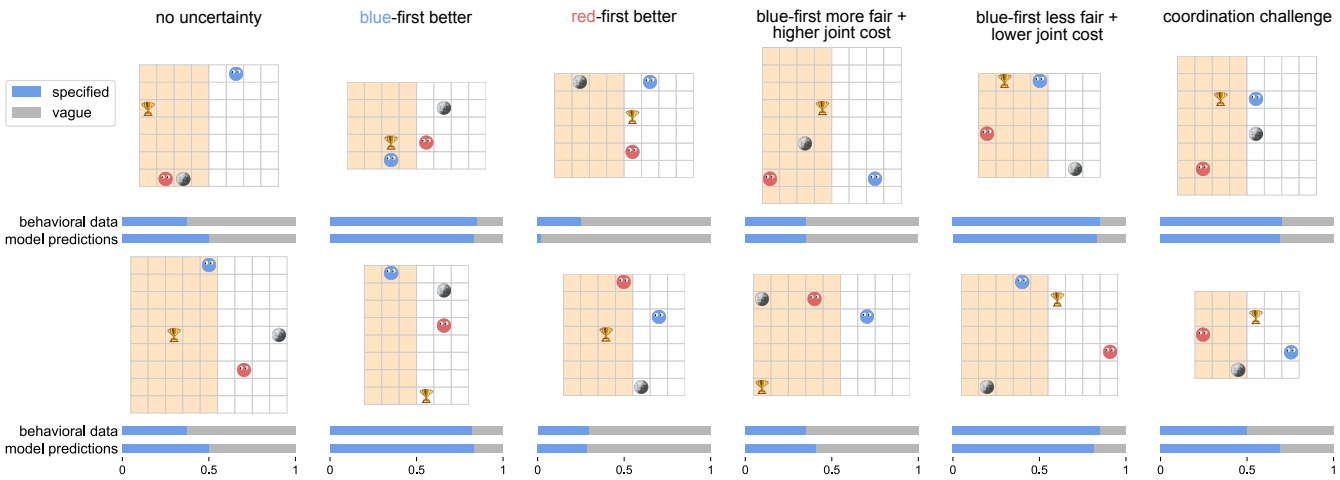

Figure 3: Behavioral data and model predictions for plan selection, across the 6 grid types.

pants should be indifferent between plan types for no uncertainty grids, because the blue-first and red-first assignments are the same, and therefore there is no risk of coordination failure; however, people exhibited a moderate preference for the vague plan. This could reflect a general aversion to longer utterances (Goodman & Frank, 2016). The model also predicted a preference for the specified plan for the second coordination challenge grid (Figure 3, column 6, row 2), yet people were evenly split. One possible explanation is that participants anticipated using perceptual information from the grid to disambiguate the vague plan (inferring that the intended assignment was the one that prescribed each player picking the item in their own uniquely-colored half).

Alternative models We assessed two alternative assumptions about how people interpret the vague plan – different schemas for computing $P(\rho|\sigma,g)$. One assumes both orderings are equally probable under the vague plan: $P(\rho_{\text{blue-first}}|\sigma,g) = P(\rho_{\text{red-first}}|\sigma,g) = \frac{1}{2}$. The other assumes people interpret the vague plan as prescribing the *alternative* assignment to the one implied by the specified plan (i.e. the red-first assignment): $P(\rho_{\text{red-first}}|\sigma,g) = 1$. This assumption interacts with the weight terms: if a participant believes that the red-first assignment is unequivocally prescribed then coordination failure is ruled out ($w_F = 0$).

Both of these models also align with the data well (mean JSDs 0.075, 0.090, respectively). However, the equal-probability parameterization tended to underestimate the likelihood of the vague plan across grids, particularly in the grids where the joint cost of the blue-first assignment was higher. The red-first only parameterization tended to overestimate the rate of the vague plan in scenarios where there is a clear reason to favor the specified plan to avoid coordination failure.

We also evaluated model variants that included a penalty for more unfair outcomes as well as a joint cost calculation that preferentially weighted the participant's own cost minimization over their partner's; we found that these alterations did not improve model fit. We do not interpret this as evidence that participants disregard these factors altogether; rather, for the specific grids we evaluated, we did not find conclusive evidence that they were strongly weighted.

Experiment 2

Experiment 1 confirmed that people make strategic and systematic use of vagueness when selecting joint plans for specific situations (grid configuration). In Experiment 2, we examined whether strategic use of vagueness extends to the setting of generic testimony, in which participants were asked to integrate the value of different plans over a structured *distribution* of grid types when advising future learners. We manipulated the distribution of grids to which participants were exposed during plan-selection trials, and assessed whether exposure to increased ambiguity systematically increased propagation of vague plans during social transmission.

Experimental protocol

The protocol was the same as the previous experiment, with 3 modifications. First, we designed new sets of grids to show participants during plan selection trials. Participants were assigned to one of 4 conditions, and selected plans for 10 grids. In the **blue-first biased condition**, participants primarily viewed grids in which the blue-first assignment is unequivocally better. There were 8 (distinct) grids where the blue-first order is better and 2 (distinct) grids where the blue-first and red-first orderings prescribe the same actions (*no uncertainty* grids). In the **unbiased** condition, grid composition was balanced: there were 2 *no uncertainty*, 4 *blue-first better* and 4 *red-first better* grids. In the **moderate red-first biased** condition, there were 2 *no uncertainty*, 2 *blue-first bette* and 6 *red-first better* grids. In the **strong red-first biased condition**, there were 2 *no uncertainty* and 8 *red-first better* grids.

Second, we modified the grids shown during the task familiarization trials. These grids may also impact participant decisions during testimony for future players; thus, the composition of grids in the task familiarization trials should be consis-

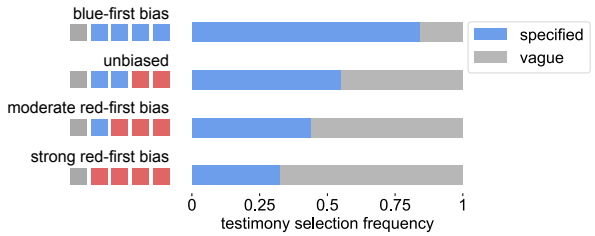

Figure 4: Selection frequency for testimony trials, across the 4 experimental conditions. Y-axis labels visually represent the frequency with which the no-uncertainty, blue-first better, and red-first better grids appeared during learning; each square represents 2 trials.

tent with the composition of grids in the plan selection trials. In the **blue-first biased** condition, participants completed 2 *no uncertainty* grids, and 2 *blue-first better* grids, identifying assignments that are consistent with all 3 plans across those 4 grids. In the **unbiased condition** and the **moderate red-first biased** condition, the composition of grids was the same as Experiment 1: 2 *no uncertainty*, 1 *blue-first better*, 1 *red-first better*. In the **strong red-first biased** condition, participants completed 2 *no uncertainty grids*, and 2 *red-first better* grids.

Finally, participants endorsed one of the two plans as testimony for new participants. People were asked to select the plan they think will be most helpful for a future naive player who will be asked to complete the same set of grids. We refer to this trial as the **testimony** trial, to contrast it with the 10 preceding **single grid plan selection** trials. Participants were told that their bonus could be linked to the performance of the future player, incentivizing them to endorse strategically. During this trial, participants could choose to endorse either the specified plan (blue-first ordering) or the vague plan.

We pre-registered[4] a target sample size of 200 participants (50 per condition). Participants were assigned at random to one of the 4 conditions. We excluded from analysis participants who failed the quiz or did not complete all the trials, yielding $N = 51, 49, 50, 49$ participants in the blue-first biased, unbiased, moderate red-first biased and strong red-first biased conditions, respectively. Median time taken was 17 minutes; participants were compensated $3.75 (median hourly rate: $13/hr).

Results

Integration of grid distribution into testimony Participants systematically integrated distributional information into their endorsement decisions during testimony (Figure 4). Participants strongly preferred to transmit the specified plan in the **blue-first biased** condition; they were more likely to transmit the vague plan in the unbiased condition ($\beta = 1.51$, CrI=$[0.6, 2.47]$), the moderate red-first biased condition ($\beta = 1.96$, CrI=$[1.05, 2.93]$), and the strong red-first biased condition ($\beta = 2.45$, CrI=$[1.52, 3.43]$). Consistent with Experiment 1, participants were more likely, compared to the blue-first better grids, to select the vague plan for no uncertainty ($\beta = 2.03$, CrI=$[1.67, 2.42]$) and red-first better ($\beta = 3.44$, CrI=$[3.05, 3.85]$) grids (data aggregated across all conditions, with a random effect for the participant).

Individual decisions predict testimony Within treatments, the number of times a participant selected the vague plan across the single grid trials predicted the plan selected during testimony ($\beta = 0.55$, CrI=$[0.24, 0.91]$). This suggests that participants who endorsed the vague plan in the testimony trial recognized an asymmetry in affordances between the two plan types: the vague plan can be interpreted as consistent with the optimal assignment in *blue-first better grids*, but the specified plan cannot be reliably interpreted to prescribe the optimal assignment in the *red-first better grids*. This result is consistent with the idea that people endorsed the vague plan during testimony because they recognized the value of vagueness during individual grid selections and integrated effectively across the distribution of their decisions. An additional possible mechanism also came to light in the **unbiased** condition: of the 22 people who endorsed the vague plan during testimony in this condition, 12 of them did so having selected the vague plan for the single grid trials 5 or fewer times (less than the majority). This raises the possibility that people may also reason post-hoc about the benefits of the vague plan.

Discussion and conclusion

We developed an experimental paradigm to study how effectively people make strategic use of specificity and vagueness during joint planning. Across two experiments participants used and endorsed vague plans selectively, in a manner consistent with a model of value-based joint planning. Our study was limited in important ways that motivate further research. Our experiment examined a simplified decision problem that removes key components of naturalistic joint planning, most notably the ability to produce open-ended natural language plans and to engage in multi-turn interactions. It is possible that asking participants to spontaneously generate plans, instead of providing them with a fixed set from which to choose, would surface different choice criteria, particularly in contexts where the provided specified plan was suboptimal. Another limitation is that our experiment did not examine a natural third contrast to our vague and specified plans: no plan, i.e. a non-communicative "virtual bargaining" solution to the problem (Misyak, Melkonyan, Zeitoun, & Chater, 2014). While it is easy to imagine situations where a vague plan is much better than no plan, it is not clear that this would be the case in simple grid games, motivating richer planning tasks in future work. Finally, our modeling framework was not designed to identify a definitive model for the structure of value computation for vague plans. Though our heuristic approach to $P(\rho|\sigma, g)$ was sufficient to make systematic predictions about the relative value of two simple plans, we anticipate many opportunities for more comprehensive investigation of these computations in future research.

[4]Link to pre-registration: https://osf.io/6vamx/?view_only=2ee407c359494d269c186cd9574b2e79

References

Acquaviva, S., Pu, Y., Kryven, M., Sechopoulos, T., Wong, C., Ecanow, G., ... Tenenbaum, J. B. (2022). Communicating Natural Programs to Humans and Machines..

D'Andrade, R. G. (1981). The Cultural Part of Cognition. *Cognitive Science*, *5*(3), 179–195.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Gordon, J., Knoblich, G., & Pezzulo, G. (2023). Strategic Task Decomposition in Joint Action. *Cognitive Science*, *47*(7), e13316.

Grice, H. P. (1975). Logic and Conversation. In D. Davidson (Ed.), *The logic of grammar* (pp. 64–75). Dickenson Pub. Co.

Hawkins, R. D., Gweon, H., & Goodman, N. D. (2021). The Division of Labor in Communication: Speakers Help Listeners Account for Asymmetries in Visual Perspective. *Cognitive Science*, *45*(3), e12926.

Ho, M. K., Saxe, R., & Cushman, F. (2022, November). Planning with Theory of Mind. *Trends in Cognitive Sciences*, *26*(11), 959–971.

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., & Tenenbaum, J. B. (2016). Coordinate to cooperate or compete:Abstract goals and joint intentions in social interaction. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *38*(0).

Lassiter, D., & Goodman, N. D. (2017). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, *194*(10), 3801–3836.

McCarthy, W. P., Hawkins, R., Wang, H., Holdaway, C., & Fan, J. E. (2021). Learning to communicate about shared procedural abstractions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Misyak, J. B., Melkonyan, T., Zeitoun, H., & Chater, N. (2014, October). Unwritten rules: virtual bargaining underpins social interaction, culture, and society. *Trends in Cognitive Sciences*, *18*(10), 512–519.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.

Stacy, S., Li, C., Zhao, M., Yun, Y., Zhao, Q., Kleiman-Weiner, M., & Gao, T. (2021). Modeling communication to coordinate perspectives in cooperation. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Sumers, T. R., Ho, M. K., Hawkins, R. D., & Griffiths, T. L. (2023, March). Show or tell? Exploring when (and why) teaching with language outperforms demonstration. *Cognition*, *232*, 105326.

Tessler, M. H., & Goodman, N. D. (2019). The language of generalization. *Psychological Review*, *126*(3), 395–436.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005, October). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(5), 675–691.

Török, G., Pomiechowska, B., Csibra, G., & Sebanz, N. (2019). Rationality in Joint Action: Maximizing Coefficiency in Coordination. *Psychological Science*.

Wu, S. A., Wang, R. E., Evans, J. A., Tenenbaum, J. B., Parkes, D. C., & Kleiman-Weiner, M. (2021). Too Many Cooks: Bayesian Inference for Coordinating Multi-Agent Collaboration. *Topics in Cognitive Science*, *13*(2), 414–432.

Zhi-Xuan, T., Ying, L., Mansinghka, V., & Tenenbaum, J. B. (2024). Pragmatic Instruction Following and Goal Assistance via Cooperative Language-Guided Inverse Planning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (pp. 2094–2103).