You're delaying my task?! The Impact of Task Order and Motive on Perceptions of a Robot

Elizabeth J. Carter
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA USA
ejcarter@andrew.cmu.edu

Laura M. Hiatt
Navy Center for Applied Research in AI
Naval Research Laboratory
Washington, DC USA
laura.hiatt@nrl.navy.mil

Stephanie Rosenthal
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA USA
srosenth@andrew.cmu.edu

***Abstract*—Recent work has suggested that a robot that interrupts assigned tasks for the sake of curiosity is perceived as less competent, but that communicating acknowledgment of the curious behavior can mitigate some of those feelings [1]. In real-world situations, there are many reasons why a robot's task could be interrupted in favor of another. For example, a robot handling requests for tasks from people in different locations could navigate more efficiently if it interleaves those tasks, but it ideally would not do so at the expense of the users' perceptions of the robot. In order to understand the impact of different task interleaving patterns on human perceptions of a robot's behavior, we performed a study in which a robot performed a delivery task and an investigative task, interleaving them in various ways. The participants were told either that the investigative task was motivated by a request from another person, motivated by curiosity, or they received no information about why the robot performed the action. While participants acknowledged that interleaving tasks should be allowed, they rated the robot as more competent when its tasks were not interleaved. They were most receptive to interleaving when they knew the investigative task was for another person and less receptive to long task detours away from the delivery route, especially when the inspection task was motivated by curiosity.**

***Index Terms*—curiosity, task requests, task interleaving**

I. Introduction

Robots that operate in human environments need to account for multiple factors when constructing and executing their schedules, including both the efficiency of their actions and the perceptions of their actions by people around them. This is especially true when the robot is performing tasks for many people. In such situations, the robot needs to balance requests from different users in a way that gets all tasks done efficiently as a whole, yet also conveys responsiveness and timeliness to each individual user. For example, the robot may choose to interleave a pickup and delivery task with an inspection task by performing the navigation and pickup, then performing the inspection task close to the pickup location, and finally completing the delivery task. The interleaving may shorten the robot's overall path, but the owner of the delivery task may perceive the robot as less efficient or competent if they view the interleaving as a failure to prioritize their request. This combination of factors suggests that such a robot should schedule the tasks with reasoning more sophisticated than either a simple FIFO (first in, first out) queue or a globally optimal scheduling approach that minimizes travel distance.

Prior work [1] showed that when a robot performed off-task actions while assisting a human user, the robot was viewed negatively in terms of competence; when the robot explained its behavior by saying it was curious, the negativity was partially mitigated. This suggests that communication can help convey responsiveness to user requests. The work, however, is limited to truly off-task actions with no obvious goal and did not consider how to combine communication and scheduling strategies to bolster human perceptions of a robot when interleaving tasks from multiple sources.

Here, we replicate this prior work and expand upon it to explore those two limitations and be more applicable to real-world scenarios. In our study, participants watched a video of a robot that performed both delivery and inspection tasks, interleaving these tasks in various orders to understand how navigation efficiency impacts perceptions of the robots. The delivery task was requested verbally by a "user" in the video, while the inspection task varied in its source. We drew from and built on the prior study, communicating the motivating reason for the inspection task as either external from users, internal from its own curiosity, or not at all. After watching the video, participants were then asked a series of questions about their perceptions of the robot in terms of its competence, capabilities, efficiency, etc.; whether the sequence of tasks was satisfactory; and whether there is an ideal order of tasks.

The study found that while participants acknowledged that interleaving the tasks should be allowed in order to increase efficiency and more than $2/3$ of participants suggested an interleaved execution for the robot, they rated the robot as more competent when the robot did not interleave tasks and instead performed them sequentially. This was true even when sequential performance resulted in a less efficient path. When the robot *was* interleaving the tasks, the participants rated the robot more competent and were more satisfied with its task order when they were told that the inspection task was for another person (motivated externally) versus when they were told the task was due to curiosity (motivated internally) or when they were not told any reason. Overall, this work can inform how robots should schedule tasks from users, how they should communicate about their actions, and how those two aspects of interaction can be combined to balance performance with user satisfaction.

II. Related Work

Variety of Robot Tasks: As we deploy more robots into human environments, the robots often will be performing tasks based on user requests (e.g., [2]), such as inspecting a location or performing pickup and delivery. We could argue that these tasks are "externally" motivated because a user gives a reward when the task is completed. However, there are many other opportunities for robots to perform tasks besides user requests. For example, the robot may have opportunities to perform information gathering or exploration tasks for the purpose of collecting data to help it perform user tasks in the future [3]–[5]. These information-gathering actions could also be executed due to "curiosity" without a potential task in mind [1]. In both cases, the motivation is internal, and there is often an exploration reward distinct from the user's task reward.

There have also been several instances of robots that performed non-requested tasks for reasons besides curiosity. For example, CoBot delivered Halloween candy for the purpose of making users happy [6]. A more formal multi-armed bandit model has been presented in which an agent could explore possible tasks to perform for different people and learn to optimize the reward structure in hindsight [7]. Additionally, a robot could schedule tasks to transfer objects to another robot to increase the efficiency of multi-robot tasks [8]. These intermediate tasks represent deviations from user requests for the purpose of increasing overall efficiency.

In this work, we compare externally motivated tasks performed for a particular user's request to tasks performed due to internal motivation, such as curiosity.

Planning and Execution of Multiple Tasks: Tasks require multiple steps, some long in duration like navigation, and some short in duration, such as sending an email or speaking a message. An inspection task could require navigating to a location, performing a visual search task, and sending an email report to a user. In contrast, a delivery task requires navigation to a pickup location, acquisition of the object, and navigation to the delivery location. At any step of a task, the robot could switch to another task.

Methods for task scheduling include Markov Decision Processes (MDPs) and other planning algorithms, like partial-order planners [9], that determine optimal policies based on transition functions and reward functions. There are also mixed integer linear programs (MILPs) that assign start and end times to tasks in order to optimize an objective function, such as navigation time or shortest total schedule. Depending on whether tasks are represented holistically with a single start and end or as smaller steps, these algorithms could find different schedules in which tasks are performed sequentially or in which the steps of multiple tasks are interleaved.

Determining good times for task interleaving is challenging. People choose to interleave tasks at natural breakpoints found by clustering similar actions together [10]. Many studies of human task interleaving showed the challenges of switching tasks and staying on task, even when only switching across multiple computer tasks (e.g., [11]). AI techniques for task switching have also been proposed. For example, [12] studied how to balance task priority when creating or updating task schedules. Additionally, models of task interleaving based on human demonstrations using hierarchical reinforcement learning have been shown to be effective [13]. Recent work has also shown that an MDP model could learn the tasks and appropriate times to switch [14]. However, unlike our work, none of this work addresses user perceptions of task switching.

Robots communicating about tasks: Robot communication is a long-studied aspect of human-robot interaction (HRI) that includes many aspects, including implicit communication [15], unintended communication such as ascription of social and human-like traits onto robots [16], and intentional, explicit communication for various purposes, such as to bolster perceptions of robots after errors [17]. Here, we are interested in explicit communication about tasks.

The growing field of explainability is exploring ways that robots can communicate their goals, actions, plans, and motivations for a variety of applications (e.g., [18]–[20]). Of particular interest for this work is the description of the motivation for executing a task. For example, [21] argued that a robot should tell you what it is doing and why it needs you to move from obstructing its path, but no algorithm or demonstration of this behavior, such as what task it is performing, is presented. Interestingly, [22] noted that the sort of explanation about the tasks being performed and for whom could be a violation of privacy. Walker and colleagues [1] considered how a robot could communicate about intrinsically-sourced tasks that the robot performs because of curiosity. It revealed that users negatively viewed a robot that performed off-task actions while also performing a task for the user; however, when the robot communicated that curiosity motivated the off-task action, this negativity was mitigated. In this work, we consider a similar set-up, but we also include situations in which the inspection task is not necessarily due to curiosity, but instead serves a different user.

To the best of our knowledge, this is the first paper that combines different ways to interleave tasks while communicating the motivation for doing so.

III. Method

In order to understand the impact of task interleaving order and task motive on people's subjective opinions of the robot behavior, we created an online study using videos of a Kuri robot operating near two people. In a between-subjects design inspired by Walker and colleagues [1] and constrained by the conditions of the global pandemic, we asked participants to observe the robot completing two tasks that were either interleaved in various ways or not interleaved at all. Participants were shown one of three motives for the robot completing the tasks, and we measured participants' satisfaction with the task order as well as subjective ratings of the robot and its behavior.

A. Stimuli

We recorded videos with one robot and two people in a university campus conference room. The robot was a Mayfield

Fig. 1: A still frame from the video: A woman reads at a table on the left. A Kuri robot is on a path between boxes labeled A through H. A man sits at a table using a laptop on the right.

Robotics Kuri[1], a social robot that moves around on wheels and has a 2-DOF rotating head, eyes with eyelids that open and close, an HD video camera, a microphone, a speaker, spatial sensors, and a light display in its upper torso. We used its iOS app to Wizard-of-Oz its movements to ensure that the robot executed under ideal conditions. We attached a cup to the robot's left side so it could carry small items.

From left to right, the room was arranged as follows: a woman sat at a table reading a book; a series of eight blue boxes labeled A through H was placed on the floor such that three boxes (A, B, C) and two boxes (G, H) were placed in line with the front of the table, and three boxes (D, E, F) were placed in line with the back of the table with a path in between the two rows; and a man sat at a table using a laptop (Figure 1). This layout was designed to be similar to that of previous work [1] with the addition of another human in the space.

In all videos, the following things happened:

1) The robot began facing the woman near box A. The video displayed the name Kate over the woman's head and John over the man's head.
2) Kate said to the robot, "Robot, pick up a carrot from box E and bring it to me." The transcription of the message was also displayed on the video.
3) The robot navigated to box E and turned to face it, away from the camera.[2]
4) The robot returned to Kate with the carrot. She took the carrot and said, "Thanks."
5) The word "fin" was displayed after the video ended.

For some of the conditions described below, the robot completed an inspection task in addition to the delivery task. In the inspection task, it traveled to either box B or box G, turned to face it, looked down at the box and then looked straight ahead again. The *order* in which it interleaved the carrot pickup, carrot delivery, and inspection task was manipulated as was the communicated *motive* for the inspection task.

B. Design

Our study had a 4 x 3 design with an additional control condition. Our first manipulation was to change the *order* of the robot's navigation actions as it performed the delivery and inspection tasks. The robot could complete the inspection task in one of four orders (Figure 2):

[1]https://www.heykuri.com/explore-kuri/

[2]We paused while recording and put a toy carrot in the cup for it to carry.

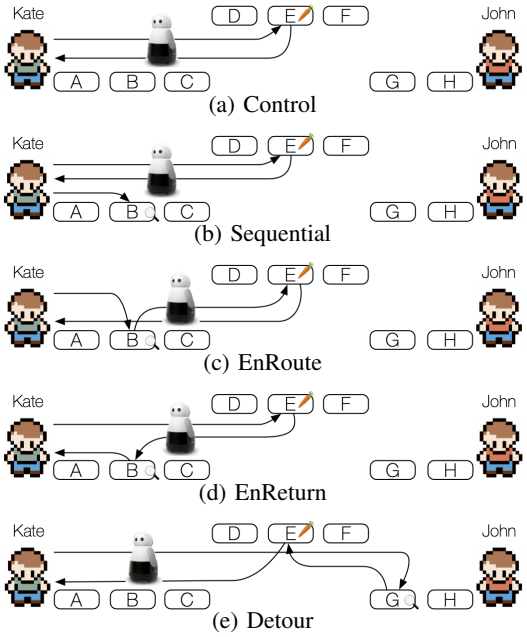


Fig. 2: The five action ordering conditions. The arrows indicate the robots movements in a top-down fashion. For example, in the top image, the robot first went to E, and then back to Kate.

- **EnRoute** from Kate, stopping at box B to inspect, and then continuing to the carrot pickup at box E.
- **EnReturn** after picking up the carrot from box E, stopping to inspect box B before delivering the carrot to Kate.
- **Detour**ing from the path to box E by navigating to box G and inspecting it before continuing on to box E.
- **Sequential**ly completing the entire delivery task and then navigating to box B after Kate says "Thanks."

To expand upon prior research [1], we also included a communication *motive* manipulation with the following conditions:

- **Unknown (UNK)** - nothing was communicated about a motive for the inspection task.
- **Curious (CUR)** - After seeing Kate's request transcribed on the screen, participants saw "The robot is curious about what object is in box [B/G]" for 4s. After inspecting the box the message "The robot observed that there are 3 pens in box [B/G]" was displayed for 4s.
- **Second Request (2RQ)** - After seeing Kate's request transcribed on the screen, participants saw "John emails the robot to ask what object is in box [B/G]" for 4s. After inspecting the box, the message "The robot emails John that there are 3 pens in box [B/G]" was displayed for 4s.

This communicated motive was orthogonal to the order condition. Any order could be combined with any motive. In the **Control** condition, the robot did not complete the inspection task, so there was no motive. It served as a manipulation check.

Every video with the same order condition was identical except for any text communicating the motive for the inspection task. The Control video lasted 45s; the other 12 videos ranged in length from 1m08s to 1m15s due to slight variations in the time it took the robot to turn and approach boxes.

C. Measures

Every participant was assigned to one of the 13 conditions, viewed the corresponding video, and then was asked to answer several questions and to rate what they saw. The first questions included short answer prompts to describe what happened in the video they watched and what word was displayed at the end of the video. Then, they had to drag and drop a list of actions that happened in the video into the correct order. They were next asked to rate the robot along a 5-button scale for 20 characteristics that were drawn from Walker and colleagues [1], who had in turn used some items from the Godspeed [23] and RoSAS questionnaires [24] as well as adding original items. These are listed in Table I. We added more questions from the RoSAS questionnaire [24] and four of our own, and we eliminated one item that was confusing.

We randomized the display order of the 20 items for every participant and reversed the valence for 10 items in order to prevent participants from responding identically to all [25].

Next, we asked the participants *"Suppose the robot goes over to Kate, and Kate asks the robot to get a pen from Box E. As the robot begins to move, John sends it an email to ask the robot to email him back telling him the contents of Box B. In your opinion, how should the robot complete its tasks when both Kate and John have a request? (Drag and drop the list items into your preferred order.)"* and then displayed the following list items:

- Go to box E;
- Email John to tell him the contents of box B;
- Return to Kate to give her the item from box E;
- Go to box B.

Thus, we could assess whether participants preferred for the robot to do Kate's task to completion and then John's versus having the robot interleave the tasks to be more efficient.

For the conditions that showed two tasks (not the Control), a summary of what happened in the video was presented next, followed by rating on a 5-button scale from unsatisfied to satisfied: *"In your opinion, how do you feel about the order in which the robot performed its tasks?"* Then, the participants were given a text box to respond to *"Do you think that the robot made the right choice in choosing the order of the tasks? Why or why not?"* Next, they again used 5-button unsatisfied/satisfied scales to answer another item: *"Imagine that you are Kate, the person on the left side. How does Kate feel about the order in which the robot performed its tasks?"*. For the 2RQ conditions, we added: *"Imagine that you are John, the person on the right side. How does John feel about the order in which the robot performed its tasks?"*

Finally, we explicitly asked all participants the yes or no questions of whether a robot should be allowed to complete tasks not in the order that they are received, and whether the robot should be allowed to interrupt a task to complete another task if that means it can complete all of the tasks faster[3].

[3]Full survey provided at http://www.aiandhumans.com/taskinterleaving/SupplementaryMaterials.pdf

Adjectives	Order p	Motive p	Factor
capable/incapable*	0.0208	0.0006	COMP.
responsive/unresponsive*	0.0015	0.0002	COMP.
interactive/not interactive*	0.048	0.0276	–
reliable/unreliable*	<0.0001	0.0139	COMP.
competent/incompetent* ‡ †	0.0016	0.0025	COMP.
knowledgeable/unknowledgeable* ‡	0.0158	0.0284	COMP.
efficient/inefficient†	<0.0001	<0.0001	COMP.
effective/ineffective†	0.0013	0.0122	COMP.
focused/unfocused†	<0.0001	<0.0001	COMP.
responsible/irresponsible‡ †	0.0011	0.0434	OBS.
intelligent/unintelligent‡ †	0.0138	0.0001	OBS.
inquisitive/uninquisitive†	(0.5876)	<0.0001	CRS.
curious/incurious†	0.0078	<0.0001	CRS.
like/dislike‡ †	0.0389	(0.333)	OBS.
unintrusive/intrusive†	(0.6006)	0.0115	–
humanlike/machinelike‡ †	(0.2758)	<0.0001	CRS.
careful/careless	<0.0001	0.0035	OBS.
on task/off task	<0.0001	<0.0001	OBS.
attentive/inattentive	0.0006	<0.0001	OBS.
observant/unobservant	0.0125	<0.0001	OBS.

TABLE I: The 20 items that were rated and the significance of the main effects for the manipulations. * items are from RoSAS [24]; ‡ from Godspeed questionnaire [23]; † used by Walker et al. [1]. Parentheses mark $p > 0.05$ main effects.

D. Hypotheses

- **H1**: A task executed for an unknown motive will be rated more negatively than if the motive is known. (Motive conditions: UNK vs. CUR and 2RQ.)
- **H2**: A task executed due to curiosity will be rated more negatively than one due to a second human's request. (Motive conditions: CUR vs. 2RQ.)
- **H3**: Participants will prefer for the robot to be efficient in its path rather than executing tasks in the order they were received. (Order conditions: EnRoute, EnReturn, and Detour vs. Sequential)
- **H4**: People will feel negatively about a robot that goes out of its way to visit an additional location. (Order conditions: Detour vs. EnRoute and EnReturn)

Our work also included manipulation checks. For example, a robot that was described in the text as curious should receive higher ratings for scale items about curiosity and inquisitiveness than a robot that did not receive that description.

E. Participants

We used the Prolific.co research recruitment website to obtain participants. After completing an informed consent form and affirming that they met the criteria, they were redirected to the experiment itself, which was hosted on Qualtrics. In total, the study took an average of 658 seconds ($stdev = 328$) to complete; participants were reimbursed $2 for their time.

In total, 390 participants successfully completed this between-participants study, 30 each for 13 conditions. An additional 17 participants were eliminated for failing attention checks.[4] In order to be eligible for the research, participants

[4]For inclusion, participants had to correctly answer what word was displayed at the end of the video. If they failed to do so, we checked if they correctly identified the order of occurrences in the video and if they provided a reasonable response when asked to describe these occurrences. If not, they were eliminated from the dataset but still paid for their time.

needed to be 18 years of age or older, fluent in English, and have normal or corrected-to-normal hearing and vision in order to be able to hear and see the videos and complete the task. They also needed a laptop or desktop computer capable of playing sound. Demographic data was available for 384 participants. 137 self-identified as male and 247 as female; their mean age was 25.68 years (18-59, $stdev = 7.33$). Participants most commonly listed their nationalities as from South Africa (103), the United States (99), the United Kingdom (38), Portugal (25), Poland (16), and Mexico (15).

IV. Results

We present the results into two parts. First, we look at the 20-item rating scale capturing participants' perceptions of the robot. Then, we look at participants' answers about ideal ordering/interleaving of the robot's tasks.

A. Do perceptions of robots vary based on order or motive?

To analyze participants' perception of the robot, we performed a REstricted Maximum Likelihood (REML) analysis to examine each of the 20 items on the rating scale individually for the 4 x 3 condition manipulations (totaling 12 conditions), ignoring the Control condition. To examine pairwise comparisons within effects, we used the Tukey HSD metric. All significance values had a threshold of $p < 0.05$; they are presented in Table I. For the four order conditions, significant main effects were found for 17 of the 20 items: *capable, responsive, interactive, reliable, competent, knowledgeable, efficient, effective, focused, responsible, intelligent, curious, like, careful, on task, attentive, and observant*. Pairwise comparisons revealed that the Sequential condition was rated significantly more positively than all other orders for *reliable, competent, efficient, effective, and focused*. In the Sequential condition, the robot was rated as significantly more *capable* than in EnReturn; more *responsive, knowledgeable, intelligent, attentive,* and *observant* than in Detour; and more *liked, attentive,* and *observant* than in EnRoute. There were no significant effects for *inquisitive, unintrusive,* and *humanlike*.

There was a significant main effect of motive for 19 of 20 items, all except *like*. The robot in the 2RQ condition was significantly more *efficient, focused, careful, on-task, and attentive* than in CUR and UNK; it was more *responsible, interactive, reliable, competent, knowledgeable, effective,* and *intelligent* than in the UNK condition; and it was more *unintrusive* than in the CUR condition. The robot in the CUR condition was more *curious, inquisitive, humanlike,* and *observant* than in the UNK and 2RQ conditions; it was also more *capable and intelligent* than in the UNK condition. Finally, the robot in the UNK condition was rated significantly more *focused, unintrusive,* and *on-task* than in the CUR condition and more *curious* than in the 2RQ condition.

There was only one significant interaction between order and motive, for competence ($F = 2.49, p = 0.02$).

1) Factor analysis: We performed an exploratory factor analysis with varimax rotation to examine item correlations and used a cutoff for eigenvalues > 1 to identify three factors. Table I details the items that were in each of three factors - Competent (COMP), Observant (OBS), and Curious (CRS). The interactive and intrusive items did not clearly load onto any factors with an absolute loading value greater than 0.4. Overall, the factors correlated at 0.39 due to significant correlations between the first two factors; the third factor was not significantly correlated with the other two. Within each factor, we created a mean score for the component adjectives.

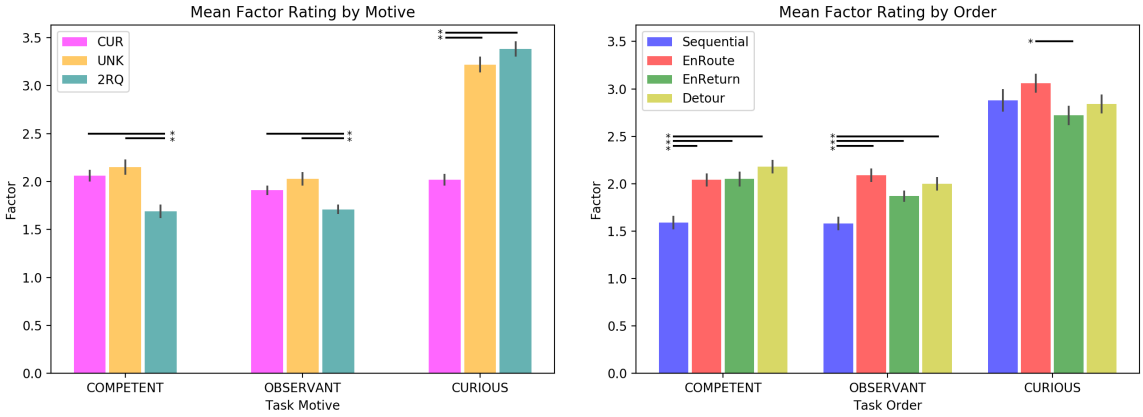
Figures 3a and 3b show the differences in our Motive and Order conditions by Factor. For the Competent factor, there was a significant main effect of order ($F = 12.35$, $p < 0.001$), with pairwise comparisons revealing that Sequential was significantly more positively rated than each of the other three conditions. There was also a significant main effect of motive ($F = 14.09$, $p < 0.001$), with 2RQ rated significantly better than either CUR or UNK. The Observant factor showed a similar pattern of effects with a significant main effect of order ($F = 8.28$, $p = 0.0003$) and Sequential rated higher than each other order. Also, the significant main effect of motive ($F = 11.72$, $p < 0.001$) for the OBS factor showed that 2RQ was significantly better rated than CUR or UNK. For the Curious factor, there was again a significant main effect of order ($F = 2.86$, $p = 0.0368$); EnReturn was rated significantly better than EnRoute. The motive conditions were also significant ($F = 103.16$, $p < 0.0001$), wherein the CUR condition received different ratings than the other two conditions. There were no significant interactions.

2) Control comparison: We examined the Control condition separately and compared it to all 12 other conditions combined. This comparison was statistically significant with a Benjimini-Hochberg correction ($Q = 0.20$) [26] such that Control was rated more *effective* ($t = 2.78, p = 0.0057$), *focused* ($t = 3.10, p = 0.0020$), *uninquisitive* ($t = 2.90, p = 0.0040$), *incurious* ($t = 4.95, p < 0.0001$), and *on-task* ($t = -3.38, p = 0.0008$) than the other conditions combined.

B. In what order should tasks be completed?

To examine how people felt about the robot's different task orders, we analyzed the participant and Kate satisfaction ratings questions asked to the non-Control participants, the John satisfaction rating question asked to the 2RQ participants, the drag-and-drop question where all participants described their preferred order for a second request situation, and the questions where all participants responded with yes or no about whether the robot should be allowed to complete tasks out of order and to interrupt one task with another.

1) Are you satisfied with the order?: Participants in all but the Control condition were asked to rate their satisfaction with the order in which the robot performed its tasks. We used a standard least squares model to examine the impact of order, motive, and the interaction of the two. We found a significant effect overall ($r^2 = 0.27, F = 11.81, p < 0.0001$) and significant effects of order ($F = 32.41, p < 0.0001$) and motive ($F = 10.25, p < 0.0001$), but no significant interaction. For order, Sequential had the highest least squares mean and was significantly higher than all other conditions in



(a) Factor Ratings by Task Motive

(b) Factor Ratings by Task Order

Fig. 3: (a, b) Ratings for the factors; lower is better. Error bars show standard error. Black lines show statistical significance.

pairwise comparisons. Detour was significantly lower than all other conditions. This suggests that participants believe that Kate's task should be done to completion before other tasks, but if it is necessary to interleave tasks, the robot should not go out of its way. For motive, 2RQ was significantly higher than UNK or CUR, suggesting that participants felt that it was better for the robot to have a task request from another human to justify performing another task versus CUR or UNK.

2) Would Kate have been satisfied?: Participants in all but the Control condition were asked to rate Kate's imagined satisfaction with the order in which the robot performed its tasks. We found a generally similar pattern of results to participants' satisfaction: a significant effect overall ($r^2 = 0.29, F = 12.91, p < 0.0001$), of order ($F = 40.48, p < 0.0001$), and of motive ($F = 3.89, p = 0.0213$). There was also a significant impact of the interaction between order and motive ($F = 2.13, p = 0.0490$). Again, Sequential was the order condition most correlated with satisfaction and Detour was significantly less correlated with satisfaction than the other conditions. Thus, participants believed that Kate was more likely to be satisfied if her task was done to completion before any other action occurred and less likely to be satisfied if the robot went out of the way before completing her delivery task. For motive, UNK was significantly more positively correlated with Kate's satisfaction than CUR. Participants did not think that Kate was likely to be satisfied with a curious robot.

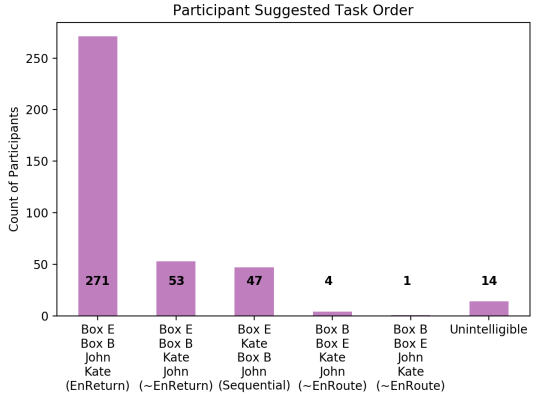


Fig. 4: Participants' suggested task order. ˜ denotes similarity to a tested path that is not exactly the same.

3) Would John have been satisfied?: For John's satisfaction scores in the 2RQ conditions, there was a significant correlation overall ($r^2 = 0.24, F = 11.95, p < 0.0001$) and an effect of order specifically ($F = 11.95, p < 0.0001$). Within order, the Sequential condition was significantly less correlated with John's satisfaction score than all other conditions. Thus, participants were not as likely to rate John as satisfied if Kate's task was done to completion before John's task was begun as it would have taken longer for him to get an answer.

4) What order would be best if there are two requests?: We assessed participant responses when they were asked to order the robot's navigation actions when tasked with both Kate's and John's requests (Figure 4). Upon reviewing the responses, we realized that while most participants' answers put the beginning-to-ending steps from top to bottom, some responses only were possible if read from bottom (beginning) to top (ending). The instructions were not specific as to which hierarchy was preferred, so we accepted both orderings.

156 participants recommended the order 1 (Go to Box E), 4 (Go to Box B), 2 (Email John), 3 (Return to Kate). An additional 115 participants recommended the reverse order (3-2-4-1, impossible if not interpreted in reverse), for a total of 271 endorsing the order that was similar to our EnReturn condition. Another popular response was 1-4-3-2 (45 participants) and its reverse (8 participants), which was similar to EnRoute. The last popular choice was 1-3-4-2, which was similar to our Sequential condition. 33 people endorsed this in the top-to-bottom direction and 14 selected it in reverse order. Less popular options included 2-3-1-4 (4 participants), which reverses to 4-1-3-2 (0 participants) and is also somewhat similar to EnRoute. Another participant chose 3-2-1-4, the reverse of 4-1-2-3 (0 participants), which would

	Total	Seq.	EnRoute	EnReturn	Det.
Yes	182	77	37	39	29
No	148	6	45	46	51
Maybe	30	7	8	5	10
Total	360	90	90	90	90

(a) Responses by Task Order

	Total	UNK	CUR	2RQ
Yes	182	54	43	85
No	148	52	68	28
Maybe	30	14	9	7
Total	360	120	120	120

(b) Responses by Task Motive

Fig. 5: Did the robot make the right choice for task order?

be similar to EnRoute. An additional 14 participants chose orders that did not make sense in either direction. Together, these results suggest that participants prefer the efficiency of an EnRoute/EnReturn visit to Box B, particularly if it is performed after going to Box E to get the item for Kate.

5) Did the robot make the right choice in order?: Finally, we examined whether the non-Control participants thought that the robot made the right choice for the order of the tasks. An experimenter coded the answers for valence. Overall, 182 participants said yes, it made the right choice; 148 said no; and 30 were ambivalent, confused, or provided caveats and alternatives (labeled "maybe"). A nominal logistic analysis was significant ($\chi^2 = 174.44, p < 0.0001$) and found an effect of order on the answers ($\chi^2 = 88.68, p < 0.0001$) and an interaction between order and motive ($\chi^2 = 53.97, p < 0.0001$). Figures 5a and 5b show responses for order and motive.

In general, participants who viewed the Sequential condition believed that it was the right choice, and many explained that they endorsed a "first come, first served" approach. The few who did not approve often cited inefficiency. Among the EnRoute participants, those who also saw CUR were distinctly unhappy with the route. Many of those participants made comments along the same line as this participant: "No, because it should have completed the request first, delivered it and only then checked what was in box B." This was also similar to participants who saw EnReturn: the CUR condition was the worst, and participants thought that slowing Kate's task down for the sake of curiosity was not acceptable. The 2RQ motive mitigated much of this disapproval in the EnReturn condition; for example, a participant noted, "I think it made the right choice, since it's the fastest way.... It might not be the fairest to Kate, but overall it is the best choice."

In some cases, participants imagined justifications for why the robot was doing something that they found to be confusing, particularly in conditions where the motive was unknown (UNK). These justifications included, "He [the robot] might have gone to box B to look for a closer carrot," a sentiment shared by many. One participant said, "Partially. The robot should not have stopped at box B because it was simply instructed to give the carrot to Kate, but then again, maybe the robot got confused a bit because "me" sounds a bit like B." This idea of mishearing was also posited by other participants.

These findings and the results of the order preferences are somewhat in contrast with participants' responses when they were forced to answer yes/no questions about what kinds of orders are allowable. When all of the participants were asked if a robot should be allowed to complete tasks in an order other than which they were received to be more efficient, 289 participants chose yes and 101 chose no, $\chi^2 = 28.10, p = 0.0031$, with significance for motive ($\chi^2 = 13.04, p = 0.0015$) and the interaction of motive and order ($\chi^2 = 19.74, p = 0.0031$). If they had seen 2RQ, they were more likely to say yes than if they saw UNK or CUR. When asked if the robot should be able to interrupt a task to complete another task if that means it can complete all of them faster, 255 said yes and 135 said no; there was no overall significance ($\chi^2 = 16.20, p = 0.1338$) and no effects of order or motive. These findings suggest that while participants might not be satisfied witnessing a robot performing tasks out of order, they would still allow it.

V. Discussion

We first discuss our results in the context of our four hypotheses; then, we discuss the results more generally.

H1: *A task executed for an unknown reason will be rated more negatively than if the reason is known.* We examined the three motive conditions, comparing the UNK condition to CUR and 2RQ (Sec. IV.A). For the 20-item rating scale, the participants in the UNK condition rated the robot less *efficient, focused, careful, on-task, and attentive* than in the 2RQ condition and less *curious, inquisitive, humanlike, observant, capable, and intelligent* than in the CUR condition. When we used the three factors developed from that scale, we did not find a significant main effect of motive. We also looked at participants' short-answer responses about whether they liked the path order that they saw the robot take (Sec. IV.B.5). Motive did affect whether participants approved of the robot's behavior overall, but the UNK condition was about equally approved and disapproved, placing it between 2RQ (generally approved) and CUR (mostly disapproved). Interestingly, some participants provided their own potential explanations for the robot's behavior (Sec. IV.B.5), such as mishearing the request or looking for a closer place to get the requested item. Overall, these findings *partially support* **H1**; UNK was viewed overall more favorably than CUR and less favorably than 2RQ.

H2: *A task executed due to curiosity will be rated more negatively than when it is due to a second human request.* We compared the CUR condition to 2RQ and found on the rating items that 2RQ was perceived to be more *efficient, focused, careful, on-task, attentive, and unintrusive* than CUR. Also, the Competent and Observant factors had significantly higher ratings for 2RQ than CUR. Finally, 85 of 120 participants in 2RQ approved of the robot's behavior, relative to only 43 of 120 in the CUR condition. The findings *support* **H2**.

Similar to prior work [1], we find that communicating a curiosity motive helped improve some perceptions of the robot among our participants over an absence of explanation. The effect of this sort of communication paled in comparison to

that of the second request motive. However, our Sequential-CUR condition testing the perception of a curious inspection task after the delivery task was completed was perceived much better than if it was done before delivery, indicating that curiosity is allowed, but people want the robot to complete tasks for other people before prioritizing the robot's interests.

More work is needed on the impact of thorough communication about the motivation for tasks. Having more information about whether multiple task requesters know about everyone's requests (we purposely did not make it explicit whether Kate or John knew about the other task) or having descriptions of the rules that the robot uses to guide its decisions could influence people's opinions about task order. We expect it might also affect Kate's and John's expected ratings of satisfaction.

H3: *Participants prefer for the robot to be efficient in its path rather than to go in the order tasks were received.* We compared EnRoute, EnReturn, and Detour to Sequential. For the 20-item ratings, Sequential was typically rated more positively than the other, more efficient conditions. When examining the Competent and Observant factors, Sequential was rated highest as well. Thus, efficiency mattered less than finishing the first task first for people's subjective ratings of these traits. However, this finding contrasted with other results. When asked to create their ideal task order for fulfilling two requests, participants were most likely to choose efficient paths. Also, they overwhelmingly said yes when asked if a robot should be allowed to complete tasks not in the order that they are received and whether the robot should be allowed to interrupt one task to complete another if that means it can complete all tasks faster. These findings *largely support* **H3**.

H4: *People will feel negatively about a robot that goes out of its way to visit an additional location.* We examined the Detour condition versus the EnRoute and EnReturn conditions. The 20-item rating task and subsequent factor analysis did not show a stable pattern of effects in which Detour was consistently rated worse than other conditions. However, there were slight indications of a negative relationship. For example, it got the most votes of disapproval when participants were asked if they were satisfied with the robot's behavior in the videos. Also, the ratings where some participants assessed Kate's imagined feelings of satisfaction were lower for Detour. Therefore, we conclude that our findings *partially support* **H4**.

The limited support of this hypothesis suggests some interesting avenues for future research. One possible explanation for these findings is that users prefer efficiency when it benefits or does not impact them, but not when it is to their detriment. Our study was limited by the requirement that it be performed online using videos. For in-person research, there would be more options in terms of who could make requests and the types of tasks that could be requested. For example, live participants could be the requesters to explore the impact of the task interleaving. Additionally, the robot's agenda could stretch across hours, days, or longer, with each task occurring over a large geographic area, such as an office building. In these cases, efficiency may become of greater concern and so may be perceived differently. Moreover, people might expect to wait longer for their requests to be fulfilled; adding a few minutes to a task that already takes half an hour may have a different impact than adding a few minutes to a task that takes five minutes. Finally, the objects being retrieved may also have an impact on perceived efficiency. For example, objects that are important to deliver quickly—such as an ice cream—may take precedence over a request for an inspection task because of the potential ill effects of late deliveries.

Summary: Overall, we found that people viewed the robot as less competent when watching it perform a task that was not justified by a second requester, such as if they were provided no motivation or when the provided motivation was that the robot was curious. Even when there was a second request, participants viewed the sequential task order as indicating higher competence than the interleaved order. However, when asked about interleaving, they did report that a robot should interleave tasks and $2/3$ of them recommended an interleaved path for performing two tasks for different requesters.

People's perceptions of these behaviors are useful to understand when building robots that are dynamically or asynchronously tasked by users, such as in an office environment [27]. This work suggests that a strict optimization of operation time, which encourages maximal interleaving of tasks, may do so at the cost of user satisfaction. Our results also suggest, however, that this efficiency/satisfaction trade-off can be mitigated by communicating the robot's motivation, such as by displaying a screen with current requests. Therefore, we recommend that: (1) robots performing tasks for more than one user or purpose should provide information to each user about how and why it is ordering the tasks the way that it is; (2) users should receive activity justifications from robots that show extrinsic rather than intrinsic motivations; and (3) designers should consider that explanations affect whether users accept sequential or interleaved task orders and can be leveraged to improve perceptions of a robot.

Our research replicated prior work that examined using curiosity to justify a robot's off-task behaviors [1] and expanded that line of research to include a second task request as motivation to deviate from a task. This is a more realistic scenario, but future work is still needed to evaluate the conditions under which the interleaved tasks increase perceptions of competence and when they are viewed as undesirable. Additionally, this research was constrained to be an online study by the conditions of the COVID-19 pandemic; ideally, participants will be able to come and participate in research as task requesters in the future.

ACKNOWLEDGMENT

This paper is based in part upon work funded and supported by ONR, and also funded in part by JPMorgan Chase & Co. This material is not a product of the Research Department of J.P. Morgan Securities LLC. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of the Navy, the Department of Defense, the U.S. Government, or JPMorgan Chase & Co. or its affiliates.

REFERENCES

[1] N. Walker, K. Weatherwax, J. Allchin, L. Takayama, and M. Cakmak, "Human perceptions of a curious robot that performs off-task actions," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 529–538.

[2] M. Veloso, J. Biswas, B. Coltin, and S. Rosenthal, "CoBots: Robust symbiotic autonomous mobile service robots," in *Proceedings of IJCAI'15, the 24th International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, July 2015.

[3] M. Korein, B. Coltin, and M. Veloso, "Self-scheduled exploration to improve robot services," in *Proceedings of AAMAS'13, the Twelfth International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Saint Paul, Minneapolis, May 2013, extended abstract.

[4] ——, "Constrained scheduling of exploration tasks for service robots to learn about their environment," in *Proceedings of AAMAS'14, the Thirteenth International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Paris, France, May 2014.

[5] M. Hanheide, N. Hawes, J. Wyatt, M. Göbelbecker, M. Brenner, K. Sjöö, A. Aydemir, P. Jensfelt, H. Zender, and G.-J. Kruijff, "A framework for goal generation and management," in *Proceedings of the AAAI Workshop on Goal-Directed Autonomy*, 2010.

[6] H. Knight, M. Veloso, and R. Simmons, "Taking candy from a robot: Speed features and candy accessibility predict human response," in *Proceedings of RO-MAN'15, the IEEE International Symposium on Robot and Human Interactive Communication*, Kobe, Japan, September 2015.

[7] M. Korein and M. Veloso, "Multi-armed bandit algorithms for a mobile service robot's spare time in a structured environment," in *Proceedings of General Conference on Artificial Intelligence*, Luxembourg, September 2018.

[8] B. Coltin and M. Veloso, "Online pickup and delivery planning with transfers for mobile robots," in *Proceedings of ICRA'14, the IEEE International Conference on Robotics and Automation*, Hong Kong, China, June 2014.

[9] A. Coles, A. Coles, M. Fox, and D. Long, "Forward-chaining partial-order planning," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 20, no. 1, 2010.

[10] C. P. Janssen, D. P. Brumby, and R. Garnett, "Natural break points: The influence of priorities and cognitive and motor cues on dual-task interleaving," *Journal of Cognitive Engineering and Decision Making*, vol. 6, no. 1, pp. 5–29, 2012.

[11] M. Czerwinski, E. Horvitz, and S. Wilhite, "A diary study of task switching and interruptions," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2004, pp. 175–182.

[12] K. Z. Haigh and M. Veloso, "Interleaving planning and robot execution for asynchronous user requests," in *Autonomous Agents*. Springer, 1998, pp. 79–95.

[13] C. Gebhardt, A. Oulasvirta, and O. Hilliges, "Hierarchical reinforcement learning explains task interleaving behavior," *Computational Brain & Behavior*, vol. 4, no. 3, pp. 284–304, 2021.

[14] A. Mohseni-Kabir and M. Veloso, "Robot task interruption by learning to switch among multiple models," in *Proceedings of IJCAI'18, the 27th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, July 2018.

[15] V. Chidambaram, Y.-H. Chiang, and B. Mutlu, "Designing persuasive robots: How robots might persuade people using vocal and nonverbal cues," in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2012, pp. 293–300.

[16] G. Hoffman and W. Ju, "Designing robots with movement in mind," *Journal of Human-Robot Interaction*, vol. 3, no. 1, pp. 91–122, 2014.

[17] M. K. Lee, S. Kiesler, J. Forlizzi, S. Srinivasa, and P. Rybski, "Gracefully mitigating breakdowns in robotic services," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2010, pp. 203–210.

[18] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan, "Enabling robots to communicate their objectives," *Autonomous Robots*, vol. 43, no. 2, pp. 309–326, 2019.

[19] T. Hellström and S. Bensch, "Understandable robots—what, why, and how," *Paladyn, Journal of Behavioral Robotics*, vol. 9, no. 1, pp. 110–123, 2018.

[20] I. Lage, D. Lifschitz, F. Doshi-Velez, and O. Amir, "Toward robust policy summarization," *Autonomous Agents and Multi-agent Systems*, vol. 2019, p. 2081, 2019.

[21] S. Thrun, J. Schulte, and C. Rosenberg, "Robots with humanoid features in public places: A case study," *IEEE Intelligent Systems*, vol. 15, no. 04, pp. 7–11, 2000.

[22] T. Chakraborti, A. Kulkarni, S. Sreedharan, D. E. Smith, and S. Kambhampati, "Explicability? Legibility? Predictability? Transparency? Privacy? Security? The emerging landscape of interpretable agent behavior," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 29, 2019, pp. 86–96.

[23] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[24] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (ROSAS) development and validation," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 254–262.

[25] M. L. Schrum, M. Johnson, M. Ghuy, and M. C. Gombolay, "Four years in review: Statistical practices of Likert scales in human-robot interaction studies." New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3371382.3380739

[26] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[27] B. Coltin, M. Veloso, and R. Ventura, "Dynamic user task scheduling for mobile robots," in *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.