Article

The International Journal of Robotics Research
1–17

DOI: 10.1177/0278364916688253
journals.sagepub.com/home/ijr

Efficient human-robot collaboration: when should a robot take initiative?

Jimmy Baraglia[1], Maya Cakmak[2], Yukie Nagai[1], Rajesh PN Rao[2] and Minoru Asada[1]

Abstract

The promise of robots assisting humans in everyday tasks has led to a variety of research questions and challenges in human-robot collaboration. Here, we address the question of whether and when a robot should take initiative during joint human-robot task execution. We designed a robotic system capable of autonomously performing table-top manipulation tasks while monitoring the environmental state. Our system is able to predict future environmental states and the robot's actions to reach them using a dynamic Bayesian network. To evaluate our system, we implemented three different initiative conditions to trigger the robot's actions. Human-initiated help gives control of the robot action timing to the user; robot-initiated reactive help triggers robot assistance when it detects that the human needs help; robot-initiated proactive help makes the robot help whenever it can. We performed a user study (N=18) to compare the trigger mechanisms in terms of quality of interaction, system performance and perceived sociality of the robot. We found that people collaborate best with a proactive robot, yielding better team fluency and high subjective ratings. However, they prefer having control of when the robot should help, rather than working with a reactive robot that only helps when needed. We also found that participants gazed at the robot's face more during the human-initiated help compared to the other conditions. This shows that asking for the robot's help may lead to a more "social" interaction, without improving the quality of interaction or the system performance.

Keywords

Human robot interaction, initiative assistive robotics, social robotics, Bayesian network

1. Introduction

In the next few decades, robots able to efficiently interact and collaborate with humans will improve productivity and the quality of everyday tasks, while reducing the workload. Robots will therefore be required to execute dexterous manipulations while fluently and efficiently performing joint tasks with humans teammates. However, many tasks, for instance in household environments, require a high level and designing robot behaviors to improve team effectiveness and fluency (Chao and Thomaz, 2013; Dragan et al., 2015). Others have given guidelines on the behavior the robot should perform in order to be perceived as social. For instance, Li et al. (Li et al., 2011) suggested that a social robot should be able to recognize the presence of human, engage in physical acknowledgment, use physical motions, express/perceive emotions and engage in some sort of communication.

of perception and cognitive and social capabilities to be efficient and positively perceived. Although such interactions come naturally to human-human teams, achieving similar fluency and comfort in human-robot teams poses many challenges.

Several of these challenges have been already tackled by previous works. Some of the main research threads have investigated ways to compute robot action plans that improve joint task performance while reducing the workload on the human (Hayes and Scassellati, 2013), tracking and anticipating human motion to enable execution of such task plans (Hoffman and Breazeal, 2010; Nikolaidis and Shah, 2013; Perez-D'Arpino and Shah, 2014),

While past work provides useful insights into *how* a robot should help as part of joint human-robot interaction and what behavior it should display to be perceived as social, in this paper we focus on the question of *when* a robot should help and how it impacts the participant's perception of the robot. In particular we investigate the factor of *initiative*

[1]Graduate School of Engineering, Department of Adaptive Machine Science, Osaka University, Japan
[2]Computer Science & Engineering, University of Washington, USA

Corresponding author:
Jimmy Baraglia, 565–0871 Asada laboratory, F1-401, 2–1, Yamadaoka, Suita city, Osaka, Japan.
Email: jimmy.baraglia@ams.eng.osaka-u.ac.jp

Fig. 1. Different initiative models for robot assistance during collaborative task executions: Human-initiated help (a), robot-initiated reactive help (b), and robot-initiated proactive help (c).

in robot assistance during *joint task execution*. We ask two questions.

1. Should the robot take initiative or let the human control the robot's participation in the task?
2. When should the robot take initiative?

To address these questions, we present findings from a user study (N=18) in which participants performed different tray preparation tasks in three conditions involving different assistance trigger mechanisms. We showed that people collaborate best with a proactive robot in terms of team fluency metrics and prefer the proactive help over other conditions. However, they preferred the human-initiated help over the reactive help, even though it results in higher human idle times and slower task completion

On top of the previous work done by Baraglia et al. (Baraglia et al., 2016), we present the autonomous system used to help a human user to achieve table top tasks. Our system uses a dynamic Bayesian network (DBN), which is able to predict future environmental states and the robot's actions to reach them. One interesting property of our model is that it can easily switch between different assistive robot. Furthermore, we show that the face gazes can be interpreted as early cues for turn taking in the human-initiated condition and could be used by the robot to speed-up its response.

The novel contributions of this article are as follows.

1. The presentation of a novel autonomous system using probabilistic inference to help a human user to achieve table top tasks.
2. A more detailed examination of the human-robot interaction subjective metrics.
3. A novel analysis of the users' gaze toward the robot during joint task collaborations and how they correlate with the previously examined metrics.

2. Related work

2.1. Human-robot collaboration

In recent years, collaborative robots designed to work side-by-side with humans have gained momentum in real-world settings. This has fueled a large body of research on human-robot collaboration. One of the core research threads tackles

behaviors by adjusting simple trigger mechanisms. Using this system, we investigate different mechanisms for triggering robot assistance in the context of joint table-top manipulation tasks. We implement three of these trigger mechanisms depicted in Figure 1:

(a) *human-initiated help*, which gives control of robot action timing to the user;
(b) *robot-initiated reactive help*, in which assistance is triggered when the robot detects that the user needs help;
(c) *robot-initiated proactive help*, in which the robot helps whenever it can.

In addition, we present additional novel analysis of the participants' gaze toward the robot's arms and face, which further supports our claims and introduces new contributions. We show that the user gazed more to the robot's face in the human-initiated condition, arguably due to the required one-way communication from a human to the

Fig. 2. (a) Goal states for the two task categories used in our evaluation. (b) Pictorial description of a sample task instance (category Task B), used to explain the task to participants in the user study.

2014) or adapting timing of motions to the human's state (Huang et al., 2015b). Chao and Thomaz (2013) developed mechanisms to coordinate sharing of common resources during collaboration, such as the speaking floor or part of the workspace that both the human and the robot need to access.

Besides generation of robot behaviors, another key problem in human-robot collaboration is perception of the human. Preliminary work by Hoffman and Breazeal suggests that anticipatory perceptual simulation improves efficiency and fluency in teamwork (Hoffman and Breazeal, 2007, 2010). With the help of new sensing and human tracking technologies, many others followed with models of action or motion anticipation in the context of human-robot collaboration (Awais and Henrich, 2012; Hawkins et al., 2014; Jarrassé et al., 2008; Nikolaidis et al., 2015).

As mentioned in Section 1, this paper focuses on the question of *initiative* about when a robot should help.

the problem of *task planning* for joint human-robot tasks. Among others, Shah generates a robot action plan so as to minimize human-idle time (Shah et al., 2011). Hayes and Scassellati developed a collaboration planner that reduces cognitive and physical load on the human (Hayes and Scassellati, 2013). Another vein of research focuses on low-level motion planning for the robot within a collaborative context (Dragan et al., 2015; Mainprice et al., 2011; Mainprice and Berenson, 2013; Sisbot et al., 2008), with an eye towards improving team fluency and the user's sense of safety.

Researchers have studied other low-level behaviors, besides robot motion, that impact collaboration and enable coordination of actions during task execution (Mutlu et al., 2013). For example, Clair and Mataric (2015) demonstrated that robot verbal feedback improves team performance. Awais and Henrich (2012) proposed mechanisms to mitigate breakdowns in joint tasks. Others focus on the coordination of micro-interactions that occur during collaboration, such as object hand-overs, using gaze (Moon et al., hesitation is detected. Similarly, Baraglia et al. (2014) proposed a developmentally motivated behavior in which the robot intervenes to help when it detects that effects of a human's action were not as predicted, i.e. the human action failed.

2.3. Initiative in human-robot interaction

In the context of human-robot collaboration, one study by Gombolay et al. (2014) is particularly relevant. They investigate decision-making authority in the planning process and find that people are willing to give control to the robot for the efficiency benefits Gombolay et al. (2014). While our results are consistent with theirs, our study differs in its focus on authority over assistance timing *during task execution*, as opposed to authority over assistance allocation *during task planning*. Groten et al. (2010) looked at shared decision making in the context of haptic collabora-

We note that joint human-robot task planning implicitly addresses the question of *when* a robot should help by producing a plan that specifies the order and timing of human and robot actions. The key difference of the scenario considered in our work is that we do not assume pre-planning of the task prior to execution. Rather, the allocation of task components occur during task execution depending on both the human's and the robot's behaviors.

2.2. Robot assistance and help

Given our emphasis on in situ, ad hoc collaboration, rather than planned collaboration, previous work on robot help is highly relevant. In fact, many of these relate to one or more of the different help behaviors studied in this paper. For example, the work of Kwon and Suh (2013) is akin to our robot-initiated proactive help. Cuntoor et al. (2012) consider human instruction as part of the collaboration, similar to our human-initiated help condition. Najmaei and Kermani's prediction-based reactive control model for collaboration (Najmaei and Kermani, 2010) is akin to our robot-initiated reactive help. Sakita et al. (2004) design different robot assistance behaviors triggered in different conditions; such as taking over when both of the human's hands are occupied or providing verbal disambiguation when user tions. Cakmak et al. (2010) investigated initiative in robot question asking. In addition, the large body of work on mixed-initiative control in the context of robot teleoperation (Fong et al., 2003) has some relevance to our work.

2.4. Gaze in human-robot interaction

In human-human interaction, cues such as body language, paralinguistic cues and gaze shift can be used to infer intention or make turn taking. In human-robot interaction as well, similar cues can be used by the robot or the human to infer each-other's target and detect the right time to take turn during collaborative tasks.

In this context, Huang et al. (2015a) presented a study in which a robot prepares a sandwich using the ingredients ordered by a human participant. By looking at the direction of the participants gaze, the robot can predict the next command in an average of 1.8 s before the verbal indication.

Similarly, Mutlu et al. (2009) and Chao and Thomaz (2010) present studies highlighting the importance of a gaze for turn taking in the context of human-robot interaction and collaboration. Mutlu et al. then showed that user could understand the robot's turn-yielding up to 99% of the times and therefore took turn accordingly 97% of the times (Mutlu et al., 2009). Conversely, Chao et al. showed that

Fig. 3. Particular instances of the tasks used in the user study: (a) Practice task, (b) to (d) three instances of Task A, and (e) to (g) three instances of Task B performed by participants in the three different conditions.

human gaze can control the robot's turn (Chao and Thomaz, 2010).

3. System

To study different help trigger mechanisms, we develop an end-to-end system for joint task execution that allows a robot to perform object manipulation actions as well as monitor the execution of the same actions by a human. In this section, we present the details of our system.

3.1. Platform

Our system is built around the PR2 robot platform (see

human), and both-allowed (middle). Task goals are represented as a conjunction of instantiated predicates; i.e the set of relations that need to be true.

Our experiments involve six specific tasks from two task categories (Tasks A and B) in slightly different domains. All tasks in the same task category have the same set of predicates in their initial state and goal descriptions. However, specific tasks differ for particular objects and locations with which the task is instantiated. Task A involves four objects to be placed in four target locations on the tray. Task B involves six objects to be arranged on two locations on the tray. The two task categories are described in Figure 2(a) and individual task instances are shown in Figure 3. For all tasks, one object is placed on the robot-only zone so that

Figure 1). PR2 has two seven degrees-of-freedom arms giving it a large workable space for tabletop manipulation tasks. Each arm has one degree-of-freedom parallel-finger gripper that can grasp objects up to a width of 8 cm. PR2's arms are passively balanced and actuated with low-power motors, making it safe to work around humans. For perception, it has a Kinect sensor attached to the head that has a high-speed pan and tilt motion. Note that most of the system was designed independently of the platform while the action execution part was designed for and with the PR2.

3.2. Domain and task representation

We focus on joint preparation tasks. This category of tasks shares many properties of tasks previously studied in the context of human-robot collaboration (e.g. circuit building (Hayes and Scassellati, 2013), lego model assembly (Sakita et al., 2004), food preparation **?**, industrial assembly (Nikolaidis and Shah, 2013)), including partially ordered action sequencing and shared physical space. More specifically, we consider food tray preparation with n objects, m tray locations and three non-overlapping table regions. Objects can be uniquely recognized and their location is represented as a 2D coordinate on the table. For each object, we also represent its relation to other objects and targets with the three predicates is-on(object), is-at(position), and is-in(region). Note that is-on(object) is inferred based on the task knowledge, while the two other predicates are detected directly through the perception module. The table is split into three regions based on who is allowed to manipulate in them. These zones, depicted in Figure 2, are: Robot-only (near robot), human-only (near participants would need the robot's assistance at least once.

Both the human and the robot are assumed to have one task-relevant action: pick-and-place(object, x, y). The x and y coordinates can be anywhere on the table, including particular tray locations or on other objects. The action is applicable for an agent (human or robot) only on objects whose current location is within the regions allowed to the agent. In our task scenarios, one object is initially placed in the robot-only region for both tasks; two objects are placed in the human-only region for Task B.

3.3. Robot perception

The robot can segment and recognize tabletop objects using the point cloud obtained from the robot's Red, Green, Blue plus Depth (RGBD) sensor. It uses the point cloud library implementation of tabletop segmentation, which detects the table plane with the Random sample consensus (RANSAC) algorithm. It then extracts a point cloud segment corresponding to each object on the table. If an object is inside or in contact with another object, they are segmented as one object with possibly multiple colors. The robot represents and recognizes objects based on their color, location on the table and size extracted from the segmented point cloud. Color is discretized into six values (red, blue, yellow, green, pink and orange) and size into three values (small, medium and large).

The robot then estimates the current environmental state as the combination of all object states in the scene. The state corresponding to each recognized object is represented by the 3-tuple (*Color*, *Size*, *Location*). The "*Location*" variable contains one or several of

Fig. 4. Model for helping robots: Recognizes the current environmental state, predicts the possible future states using a dynamic Bayesian network and generates actions to achieve the desired end-states.

the predicates is-on(object), is-at(position) and is-in(region) presented in Section 3.2. For instance, if a "*small red cup*" at the location l_1 and a "*medium blue plate*" in the "human-only" region are recognized in the scene, the object states are noted s_1 $=(\ Red, Small, is\text{-}at(\ l_1)\ )$. and $s_2 = (\ Blue, Medium, is\text{-}in(\ human\text{-}only))$ In the case the "*small red cup*" is *on* the "*medium blue plate*" at the location l_1, the corresponding state can be noted $s_3 = (\ Red, Small, is\text{-}on(\ Blue, Medium, is\text{-}in(\ l_1))$. In practice however, when an object is *on* or *in* another object, they are detected as one new object. In this case, the object state would then be noted $s_3 = (\ \{Red, Blue\}, Medium, is\text{-}in(\ l_1)\ )$.

3.4. Robot actions

The robot's pick-and-place actions are parametrized with an *object* to be picked and a *location* at which the object is to be placed. The actions are defined as a sequence of poses relative to the object (pre-grasp, grasp, and lift poses) followed by poses relative to the target location (transfer, lower, and drop poses). While the overall action templates remain the same, some of the poses in the actions are tuned to the particular object being manipulated. The actions were trained using a learning by demonstration approach developed by Alexandrova et al. (2014).

3.5. Joint task execution model

The joint task execution model is built based on previous research in which it was showed that instrumental helping could be generated using a low level motivation signal (Baraglia et al., 2014, 2015). This work supposes that due to strong self-other correspondence, referred to as the "like-me hypothesis" (Meltzoff, 2007), the robot can project its own task state onto others performing similar acts. This mechanism allows our system in the reactive and proactive help conditions to assist users in achieving their tasks, without the need for high level trigger signals.

Fig. 5. Two time-slice dynamic Bayesian network used in this study. It is composed of two multinomial nodes S and A representing environment states and actions. The grayed node $S(\ t)$ represents the observable state.

The overall system for joint task execution is illustrated in Figure 4. At the core of this system are two modules for:

(a) tracking the state of the task and anticipating future actions;
(b) selecting a robot action based on the observed and anticipated states accreting to different help strategies.

More detailed descriptions of these modules are given in the following sections.

3.5.1. Tasks state prediction module. Our system uses dynamic Bayesian networks (DBNs) to predict future states and the robot's actions that lead to those states. DBNs are multi-time-slice Bayesian networks where variables are connected to one another over adjacent time steps as well as within the same time step. They are a computationally efficient generalization of hidden Markov models and have been used to model multi-modal robot behavior in uncertain environments (e.g. work by Huang and Mutlu (2014)).

For this study, we used two time-slices DBN. Each time-slice of the DBN contains an object state and an action node, corresponding to two multinomial discrete variables S and A. The used DBN architecture is illustrated in Figure 5. S can be one of all possible states $\{s_0, s_1, \ldots, s_N\}$ that are distinct according to the defined predicates for a finite set of objects and named locations (Section 3.2). Two states in which an object's position is different, but both positions are not at a named location, are considered the same discrete state. The variable A is one of all possible action instances $\{a_0, a_1, \ldots, a_M\}$ that involve the combination of all objects and named locations in the environment, regardless of whether they are available to the human or the robot. $S(\ t)$ represents the current observed object states in the scene, and $S(\ t+1)$ the predicted states at time $t+1$.

Fig. 6. Examples of task knowledge represented as states transition. Object $o1$ can be one of two states: $s11$ or $s12$ if positioned in $l1$ or $l2$, respectively. Object $o2$ can be one of two states: $s21$ or

Within a single time-slice, the state influences the action. Between consecutive time-slices, the state and action from the previous time-slice influence the next state.

The DBN encodes the task knowledge in the conditional probabilities $P(A(t)|S(t))$ and $P(S(t+1)|S(t), A(t))$, which represent the action policies the robot could use if it were to execute the task on its own. Since the tasks are known a priori in our scenario, these conditional probabilities were computed based on the known task structure (Section 3.2), assuming each path for completing the task is equally likely. To estimate the conditional probabilities $P(S(t+1)|S(t), A(t))$, we use a maximum likelihood parameter algorithm from a set of pre-defined data. The data are in the form of a list of state-action transitions, such as $S(t) = s_0 \rightarrow A(t) = a_0 \rightarrow S(t+1) = s_1$. Future states and actions are predicted by computing the marginal probabilities $P(S(t+1))$ using Bayesian inference. The action $A(t)$ to perform in order to transit from $S(t)$ to $S(t+1)$ is inferred by maximizing the conditional probability $P(S(t+1)|S(t), A(t))$ given a known $S(t+1)$. The result of the predictions are then sent to the action selection module. This approach is efficient and works well with fairly complex tasks. However, it may not be suitable for big state-space as the number of operations to perform the Bayesian inference is quadratic ($O(n^2)$). Improving the task state prediction module framework could help reducing the calculation time. For instance, a factored Markov decision process (Degris and Sigaud, 2010) could reduce the amount of conditional probabilities to calculate for each robot's decision making.

To illustrate these mechanisms, let us imagine a task in which a table containing two objects should be cleaned (see Figure 6). The robot's task knowledge, known a priori, contains the necessary information to represent a task. The two objects are "*small red cup*", noted o_1, and a "*medium blue plate*", noted o_2. The table is separated in two discrete locations: l_1 = "*dirtyZone*" and l_2 = "*cleanZone*". The possible object states in this example are:
s_{11} =(*red*, *small*, *is-in*(l_1)),
s_{12} =(*red*, *small*, *is-in*(l_2)),
s_{21} =(*blue*, *medium*, *is-in*(l_1)),
s_{22} =(*blue*, *medium*, *is-in*(l_2)).
The initial environmental state contains the two object states s_{11} and s_{21}. The robot can perform pick and place actions, noted a_1 and a_2, to move the "*small red cup*" or the "*medium blue plate*", respectively, from l_1 to l_2.

The transitions between the different object states as described in the task knowledge are illustrated in Figure 6. When s_{11} and s_{21} are initially recognized by the robot, marginal probabilities $P(S(t+1))$ are calculated individually for each object state. As we assume all paths for completing the task are equally likely for this task, we obtain

$P(S(t+1) = s_{11}) = 0$, $P(S(t+1) = s_{12}) = 0.5$,
$P(S(t+1) = s_{21}) = 0$ and $P(S(t+1) = s_{22}) = 0.5$

The system then infers what actions to perform in order to achieve the predicted environmental states by maximizing the conditional probabilities $P(S(t+1)|S(t), A(t))$, which are in this case equal to

$P(S(t+1) = s_{12}|S(t) = s_{11}, A(t) = a_1) = 1$
$P(S(t+1) = s_{12}|S(t) = s_{11}, A(t) = a_2) = 0$
$P(S(t+1) = s_{22}|S(t) = s_{21}, A(t) = a_1) = 0$
$P(S(t+1) = s_{22}|S(t) = s_{21}, A(t) = a_2) = 1$

Here also, conditional probabilities are calculated individually for each object state.

In this example, the robot estimates that it can perform a_1 or a_2 in order to reach s_{12} or s_{22}, respectively, from the currently observed object states s_{11} and s_{21}. When a new state is reached, the robot reiterates the same inference process until it can no longer predict new states. When reaching $S = s_{12}, s_{22}$, the system cannot predict future states based on its task knowledge, and therefore considered the current state as an end-state (or absorbing states).

When to perform an action and which action to execute is decided by the action selection module presented in the next section.

3.5.2. Action selection module. The action selection module implements a policy that specifies what the robot should do at each time step. If the robot were to execute the task completely on its own, this module would directly return one of the possible actions predicted by the DBN immediately after every action. During joint task execution, on the other hand, the robot's policy needs to account for the

Fig. 7. Examples of object detection likelihood estimation. The value increases if the object is recognized and decreases when it is not. If the object detection likelihood decreases bellow the detection threshold, the object is lost.

human's direct input or their actions that result in changes in the world state. We implement three policies that differ in terms of *when* a robot action is triggered.

1. **Human-initiated help (H):** The first policy gives complete control of robot actions to the user. The robot performs an action only when the user explicitly says "Robot, can you help me?'
2. **Robot-initiated reactive help (R):** In the second policy, robot actions are initiated by the robot when it detects that help is needed. The robot tries to detect when one of the next states predicted by the DBN is not reached within an expected time window, indicating a delay or user difficulty in the task progress.
3. **Robot-initiated proactive help (P):** The third policy involves performing actions whenever they are possible. However, different from a robot-only task execution, the robot needs to take into account human actions that might be *in progress* before a stable environmental state is reached. As the robot is not equipped with the ability to detect human actions, this is done indirectly by looking at whether or not the observed object's states are stable. If at least one executable action exists that does not conflict with human actions, the trigger is initiated.

Mechanisms behind the robot's behavior for each policy are similar, but differ in some fundamental aspects. When the system observes object states $S(t)$, it predicts future object states $S(t+1)$ that have non-null marginal probabilities as shown in Section 3.5.1. Object states are noted s_{ij}, where i is the object number and j represent the objects locations. Actions on an object i are noted a_i. To decide which action should be performed by the robot, several values are estimated.

1. Firstly, for each possible future object state, an *object detection likelihood*, noted $Ls_{ij}(t+1)$, is calculated. The value is initialized at 0.6 and increases linearly while the object is recognized (max. 1). This value represents how well an object corresponding to a predicted state is recognized by the perception module. If the object is momentarily not perceived, $Ls_{ij}(t+1)$ decreases linearly. If $Ls_{ij}(t+1)$ becomes lower than 0.4, the object is considered lost. For instance, if a user repeatedly touches an object in the scene, the corresponding $Ls_{ij}(t+1)$ will be low because the object recognition will be noisy. Examples of object detection likelihoods for good and bad object recognitions are represented in Figure 7.
2. Secondly, a *trigger signal*, noted $T\,s_{ij}(t+1)$, is estimated for each possible future object state. The value of $T\,s_{ij}(t+1)$ is a function of $Ls_{ij}(t+1)$ and of the elapsed time, noted $\bar{t}$, since the current environmental state $S(t)$ has been first recognized.

A trigger signal $T\,s_{ij}(t+1)$ is activated when $Ls_{ij}(t+1)$ gets higher than a threshold (θ) fixed at 0.8. When a trigger signal is activated, the robot executes the corresponding action a_i. The starting value of $Ls_{ij}(t+1)$ and its different thresholds were chosen empirically to ensure that the objects are well recognized before triggering a signal, and are detected even with noise and brief occlusions.

The trigger signals are calculated as follows:

1. In condition H, when a user asks the robot for help, the trigger signal with the highest object detection likelihood $Ls_{ij}(t+1)$ value is activated. If two or more trigger signals have the same object detection likelihoods, the trigger signal with the action on the closest object to the robot is activated. The distance between the robot and an object is noted d_i.
2. In condition R, when possible future states are predicted, the trigger signal values are calculated as function of the elapsed time $\bar{t}$ and the corresponding object detection likelihood as follows

$$T\,s_{ij}(t+1) = 0.3 \times Ls_{ij}(t+1) \times \frac{(\bar{t} - T)}{T} \quad (1)$$

where T represents the action duration and defines how quickly the trigger signals increase, and therefore corresponds to the robot reaction time in the reactive condition. It was here fixed at 4 seconds empirically so that the robot reaction time is neither too fast or too slow. If one trigger signal gets higher than the threshold θ, it is activated. If two or more trigger signals are higher than θ at the same time, the trigger signal with the lowest d_i is activated.
3. In condition P, when possible future states are predicted, the trigger signal values are calculated as function of the

corresponding object detection likelihood only

$$T\,s_{ij}(t+1) = Ls_{ij}(t+1) \quad (2)$$

If one trigger signal gets higher than θ, it is activated. If two or more trigger signals are higher than θ at the same time, the trigger signal with the lowest d_i is activated.

The robot always uses the gripper closest to the object of the executed action. In addition, if two or more trigger signals are activated at the same time, the one corresponding to the closest object is always preferred.

Let us now consider the example where our system and a user jointly collaborate during the task presented in Section 3.5.1 (see Figure 6). At first, the robot detects the current environmental state, noted $S(t)$. The tasks state prediction module then predicts possible future object states $S(t+1) = s_{12}$ or $S(t+1) = s_{22}$. The action to reach s_{12} is estimated to be a_1. The action to reach s_{22} is estimated to be a_2. When the states are predicted, the robot estimates for each of them an object detection likelihood value $Ls_{12}(t+1)$ and $Ls_{22}(t+1)$.

In all conditions, if the next state is correctly reached by the robot or by the user, the tasks state prediction module predicts new future states. In this example, if s_{12} is reached, the new predicted states will be $S(t+1) = s_{22}$. Conversely, if s_{22} is reached, the new predicted states will be $S(t+1) = s_{12}$.

In the H condition, the robot will not perform any of the actions until it receives a command. In the R condition, if the predicted states are not achieved within a few seconds, the trigger signal values will increase. If the user performs an action before any of the trigger signal values reach the threshold, the robot does nothing. Else, if one of the trigger signal value reaches the threshold, the robot performs the corresponding action. Finally, in the P condition the robot performs an action as soon as possible, namely when one of the object detection likelihood corresponding to predicted states is higher than the threshold.

4. User study

The help trigger mechanisms described in Section 3.5.2 are expected to yield different joint task execution dynamics. Furthermore, each mechanism on its own can result in a wide variety of behaviors depending on the particular user. For example, when interacting with the *human-initiated* policy, users may request help at every step or only when they need it. When interacting with the *robot-initiated proactive* policy, they might select their own actions such that the robot has many opportunities to help or they might (unintentionally or intentionally) block the robot's actions. The differences across and within each policy can reflect on objective task execution measures, as well as the user's subjective attitude towards the robot. To investigate these differences, we performed a user study that allows us to:

(a) characterize people's behaviors while interacting with each policy;

(b) compare the alternative policies for triggering robot help.

4.1. Study design

We performed a within participants study with one independent variable (robot helping behavior) with three conditions: H, R, P (Section 3.5.2). In each condition, participants performed two tasks with the robot, one from each category (Task A and B). The order of the three conditions were counterbalanced.

4.2. Study setup

The robot was placed in front of a 68 cm high table. Participants sat across the table. The table top was separated into three zones as shown in Figure 3. Participants were asked not to touch objects that are in the red zone (near the robot). Similarly the robot could not enter the blue zone (near the human). Both were allowed to manipulate objects in the middle zone. In the middle of the table there was a tray with four target positions.

Tasks were explained to participants with a one page pictorial description involving:

(a) the set of objects and targets involved in the task;
(b) the final state of the tray when the task is complete.

An example task description is shown in Figure 2(b). An additional small table was placed to the right of the participant. Printed task descriptions were placed on this table, together with a tablet for logging task steps (see Section 4.3) and a laptop for responding to our questionnaire. The complete setup can be seen in Figure 1.

4.3. Procedure

Participants were recruited from a campus and nearby neighborhoods through mailing lists. Interested individuals signed up for a 45 minutes time slot in advance. When participants arrived at their scheduled study time, we first explained the purpose of the study and asked them to sign a consent form. Then they were taken to the participants seat, introduced to the robot and the workspace, and given an overview of the procedure.

Next, the robot was activated and participants performed a practice task (Figure 3(a)). The task was explained to them using the corresponding pictorial description. The robot made a specific sound to indicate that it was ready. Participants were told that they can start the task when they hear this sound. They were told to perform one step of the task and then log the step on the tablet. The logging was done throughout the study as a mechanism to space human actions apart and give the robot an opportunity to detect intermediate states of the task. Each log required indicating who performed the step (human or robot), the two letter identifier for the object involved (as indicated in

the task description), and the one letter identifier for the target position where the object was placed. The second step of the task was performed by the robot to familiarize participants with the robot's motion. The robot made another sound when it detected the task completion. Participants were told that they will perform similar tasks together with the robot in three conditions where the robot's behavior will be different.

Next we moved on to the actual study. For each condition, the experimenter first gave condition specific instructions. In the human-initiated help (H) condition, participants were told that they can request the robot's help by saying "Robot, can you help me?". This was done in a wizard of Oz fashion and without using a microphone. As soon as users asked for help, the experimenter discreetly pressed a button. In the other conditions (R and P), they were told that the robot will decide when and how to help out with the task. Then the experimenter set up the initial state of the first task, told participants to start when they hear the robot sound, and left them alone with the robot. The experimenter came back to set up the next task after the robot detected that the task was complete. After completing both tasks in the same condition, participants were asked to respond to the condition-specific questionnaire. After all three conditions were complete, participants responded to additional questions drawing comparisons between the three conditions. At the end, participants were thanked for participating and given the promised compensation of a 10 USD equivalent gift card.

4.4. Measurements

The study was recorded from two cameras; one mounted on the robot's head and another overseeing the workspace together with the robot and the participant. In addition, we logged the progression of tasks and robot actions with timestamps throughout the study. The extracted data was used to evaluate three main components: The social aspect of the interaction, the quality of interaction and the system performance.

From the study logs we extracted the task completion time and the number of actions performed by each agent. From the videos we extracted quantitative measure that characterized each participant and the robot behaviors. These measures included times when the robot and the human were moving alone or in concurrence, their idle times and the number of gazes the participants performed to the face of the arms of the robot during the joint task execution. The coding was performed by two coders (IRR *= 0.72), including one without prior knowledge of the study.

To compare the three conditions subjectively from the asked them to describe their strategy. Then we asked a set of Likert scale questions, similar to those commonly used in human-robot collaboration research (Hoffman, 2013). These questions addressed the user's perception of: The robot's helpfulness, its awareness of the human and task progress, its contribution to the task, team fluency and efficiency and naturalness of the interaction (see questions in Figure 13). Additional questions at the end asked a forced ranking of the three conditions and open ended questions about perceived distinction between the two robot-initiated conditions and how different behaviors would be combined in an ideal interaction.

5. Findings

Our study was completed by 18 participants (nine females and nine males aged 18 to 35). This section presents our findings based on data collected from these participants. A repeated-measure Analysis of variance (ANOVA) was conducted to compare the effect of conditions H, R and P within subjects on the different objective metrics. We used an alpha value of $\alpha = 0.05$, which set the F-critic at $F_{crit} = 3.26$ (see F-table for $df_{Between} = 2$ and $df_{Error} = 34$). We performed post-hoc tests (two-tailed paired-t-test) to explore differences between pairs of conditions.

To analyze the experimental data, we segmented each interaction between a participant and the robot into temporal action sequences. These actions could be of three types for the participants: Acting (upon the table), logging or idling; and two types for the robot: Acting or idling. Based on these temporal action sequences, we could also extract concurrent actions between the participants and the robot. Additionally, we segmented the different gazing patterns of the participants to the robot's face and arms. Two examples of interactions for each of the three condition during the Task B are shown in Figure 8(a) to (f).

5.1. Objective metrics

We first examine common task and collaboration metrics. Figure 9(a) shows the average number of task actions performed by the robot in each condition (Task A: $F(2, 34) = 20.95, p < .05$; Task B: $F(2, 34) = 20.65, p < .05$) and Figure 9(b) shows the overall task completion times by the human-robot team (Task A: $F(2, 34) = 3.63, p < .05$; Task B: $F(2, 34) = 6.68, p < .05$).

Figure 10(a) to (d) shows the breakdown of task completion times into robot-only, human-only, concurrent, and no motion segments and Figure 10(e) to (f) separately show the human idle time and robot idle time. The results of the ANOVA for results in Figure 10 are as follows:

(a) (Task A:$F(2, 34) = 9.41, p < .05$; Task B: $F(2, 34) =$

user's perspective, we administered several questions after each condition as well as at the end. First we asked an open ended question to elicit the participants own description of the robot's assistance behavior. Another question

(b) (Task A: $F(2, 34) = 12.26, p < .05$; Task B: $F(2, 34) = 1.07, p > .05$);

. ∗ Cohen's kappa.

Fig. 8. Examples of interactions.

(c) (Task A: $F(2, 34) = 3.55, p < .05$; Task B: $F(2, 34) = 3.09, p > .05$);
(d) (Task A: $F(2, 34) = 6.23, p < .05$; Task B: $F(2, 34) = 10.71, p < .05$);
(e) (Task A: $F(2, 34) = 3.73, p < .05$; Task B: $F(2, 34) = 14.09, p < .05$);
(f) (Task A: $F(2, 34) = 6.79, p < .05$; Task B: $F(2, 34) = 7.38, p < .05$).

Finally, Figure 12(a) to (d) shows the number of times the participant looked at the robot's face and arms and the average duration of the gazes. The results of the ANOVA for results in Figure 12 are as follows:

(a) (Task A: $F(2, 34) = 3.81, p < .05$; Task B: $F(2, 34) = 10.83, p < .05$);

results in the robot having a greater contribution to the task, as indicated by the significantly higher number of actions performed by the robot (Task A: $p < .001$, Task B: $p < .001$) (Figure 9(a)). This is also reflected in the significantly lower robot idle times for the proactive robot (P) as compared to the reactive robot (R) (Task A: $p < .001$, Task B: $p < .05$) (Figure 10(f)). The average number of actions performed by the reactive robot was around 1 (Task A: $M = 1.17$, $SD = .38$, Task B: $M = 1.56$, $SD = .76$), which is the minimum number of actions required by the robot. Whereas, the proactive robot performed around 2 (Task A) and 3 (Task B) actions (Task A: $M = 2.17$, $SD = .48$, Task B: $M = 3.00$, $SD = .82$), which are about half of the actions needed to complete the task. This finding is expected and confirms that our model produced the intended behavior.

(b) (Task A: $F(2, 34) = 12.30, p < .05$; Task B: $F(2, 34) = 7.53, p < .05$);
(c) (Task A: $F(2, 34) = 12.49, p < .05$; Task B: $F(2, 34) = 2.67, p > .05$);
(d) (Task A: $F(2, 34) = 1.91, p > .05$; Task B: $F(2, 34) = 6.64, p < .05$).

5.1.1. Proactive versus reactive: First we focus on the comparison of robot-initiated help strategies. Proactive help to be greatly influenced by the difference in human and robot action speeds as humans are several orders of magnitude faster at pick-and-place actions. Hence they were

Despite the difference in the number of robot actions, there was no significant difference in the total task duration in Task A (Task A: $p = .12$) and little difference in Task B. A potential reason for this could be lack of parallelization between human and robot actions. However, the significant increase in the *concurrent* human-robot motion (Figure 10(c)) in the proactive condition indicates that parallelization did indeed happen at least in Task A (Task A: $p < .005$). In addition, the total task duration appeared

Fig. 9. (a) Number of actions performed by the robot for each task category in each condition. (b) Task completion time for each task category in each condition. Error bars represent standard deviation.

Fig. 10. Breakdown of task completion times into (a) robot-only, (b) human-only, (b) concurrent and (d) no motion time segments. These include only motion related to the joint task. (e) Human idle time. This excludes the time during which the human is performing their secondary task of *logging* task actions. (f) Robot idle time.

(H-R - Task A: $p < .05$) (Figure 10(c)). We believe that it is because participants asked for help and then started doing their own actions as soon as they understood the robot's

not slower in completing the overall task in the reactive condition. Despite this difference, human idle times were not significantly higher in the proactive robot condition (P) (Figure 10(e)).

5.1.2. Human-initiated versus robot-initiated. Next, we look at comparisons between the human-initiated help (H) condition and robot-initiated help conditions to characterize how people chose to get help from the robot when they had control. From Figure 9(a), we see that the number of actions performed by the robot in the H condition was about half of all task actions, as in the P condition. The number of actions performed by the robot was significantly higher than in the R condition (H-R - Task A: $p < .001$, Task B: $p < .001$). It resulted in significantly higher concurrent motions in Task A for the H condition compared to the R conditions

intention. This is similar to the P condition, where participants briefly waited until they recognized what the robot was doing and then acted. The added waiting time in the H condition was reflected in overall task completion times (Figure 9(b)), which was significantly higher than in the R condition for Task B (H-R - Task B: $p < .005$) and in the P condition for both tasks (H-P - Task A: $p < .05$, Task B: $p < .05$). This was also reflected in the human idle times (Figure 10(e)) which was highest for the H condition in both tasks (H-R - Task A: $p = .27$, Task B: $p < .001$; H-P - Task A: $p < .05$, Task B: $p < .01$). We noticed that one participant made the robot do all actions for Task 1; two participants made the robot do all possible actions for Task 2 in the H condition. This contributed to the high human idle time and task completion time, while making the variance in this condition high.

Fig. 12. (a) Average number of gaze to the face of the robot; (b) average number of gaze to the robot's arms; (c) average duration of each gaze to the face of the robot; (d) average duration of each gaze to the robot's arms.

$p < .001$) (see Figure 12(a)). The average gaze duration was also significantly longer in the H condition during Task A (H-R - Task A: $p < .005$; H-P - Task A: $p < .005$) (see Figure 12(c)). This can be explained by the participants having to vocally command the robot when needing the robot's help. In fact, the participants almost always gazed to the robot's face when asking for help and kept gazing until the robot would start its action.

The number of gazes to the arms of the robot is significantly greater for the H and P conditions compared to the R

Fig. 11. Gazing zones: The face gaze zone (purple line) is situated on the robot's "head" part. The arms gaze zone (red line) is situated on the lower body part of the robot.

5.1.3. Gaze. We then look at the participants' gazing patterns toward the robot during the different tasks. Two gazing targets were analyzed: The robot's face and arms zones, which are shown in Figure 11. The number of times the participants looked at each zone and the duration of each gaze were extracted from the video recording of the experiment. The gazes to the robot's arms were only counted when the robot was moving its arms.

The number of gazes to the face and arms of the robot are described in Figure 12(a) and (b), respectively. The average duration of each gaze to the face and arms of the robot is described in Figure 12(c) and (d), respectively.

The participants gazed significantly more to the face of the robot in the H condition compared to the R and P conditions during Task B (H-R - Task B: $p < .05$; H-P - Task B: $p < .05$; Task B: $P_{coef} = 0.31, p > .05$). In the case of face gazes, this can be explained by the fact that users looked at the robot's face after asking for help and kept looking until the robot started moving. However, users only looked at the robot's arms when it was moving, which explains the absence of correlation.

In addition, we found correlations between gazes to the face and to the arms and some objective metrics. In studies by Huang et al. (2015a) and Mutlu et al. (2009), it was shown that gazes could be strong indicators for turn taking or intention prediction. As our experiment involves a human and robot performing joint tasks, turn taking is an intrinsic part of the interaction. By observing the cues given by the users' gaze, the robot could for instance predict when they need help and be more efficient at helping. To find out if such cues were given by the participants, we looked at the participants' gazes 5 to 15 seconds prior the onset on the robots' action. In the human-initiated condition, 70.56% of all face gazes preceded a robot action (Task A: 76.67%; Task B: 64.44%). In the robot-initiated reactive help, this value is 24.96% (Task A: 23.81%; Task B: 25.93%). In the robot-initiated proactive condition, it is 17.23% (Task A: 10.64%; Task B: 23.81%). These results indicate that users gaze to the face of the robot more when they need to ask the robot for help than when the robot acts by itself. However, they also receive slightly more gaze when the robot was reactive rather than proactive. It shows that the more condition in Task A (H-R - Task A: $p < .05$; R-P - Task A: $p < .001$) and greater in condition P compared to the R condition in Task B (R-P - Task B: $p < .001$) (see Figure 12(b)). This result is strongly correlated with the number of actions performed by the robot (see Figure 9(a)). We argue here that the amount of gaze to the arms is an artifact of the logging requirement. Indeed, gazing to the robot's arms allows the participants to identify its actions and to log them as part of the task. Therefore, if the robot executes more actions, the number of gazes should increase proportionally.

Next, we found that the average gaze duration to the robot's face and arms was significantly longer in the H condition compared to the two others in Task B (H-R - Task B: $p < .05$; H-P - Task B: $p < .05$) (see Figure 12(d)). The average gaze duration was correlated to the tasks average completion time (see Figure 9(b)) with a Pearson product-moment correlation coefficient (hereafter noted P_{coef}). We found significant correlation between the average completion time and face gazes, but not with arm gazes: (Cond H - Task A: $P_{coef} = 0.57, p < .05$; Task B: $P_{coef} = 0.60, p < .05$; Cond R - Task A: $P_{coef} = 0.50, p < .05$; Task B: $P_{coef} = 0.42, p > .05$; Cond P - Task A: $P_{coef} = 0.76, P_{coef} = 0.65, p < .05$). These results give us the following insights on the interaction.

1. Users looked more at the face of the robot when it acted alone ("robot only" and "human idle"), showing signs that users tried to understand the robot's intention.
2. Users also gazed more at the face of the robot when it was not moving ("no motions" and "robot idle"), arguably because users were waiting for the robot to take its turn as shown by Chao and Thomaz (2010) and Mutlu et al. (2009).
3. Users gazed more at the robot's arms when it was moving alone and when users were logging ("robot only"), as already observed in the results presented in the previous paragraph.

5.2. Subjective metrics

Participant responses to the Likert-scale questions are summarized in Figure 13. The inter-condition differences were analyzed using the Wilcoxon signed rank test (we also conducted parametric tests and obtained similar results). which is a standardly used non-parametric test. As suggested in the work by Carifio and Perla (2007) and to avoid family-wise errors, we grouped the seven scales into two sub-scales representing the quality of interaction (Figure 13(a)) and the system performance (Figure 13(b)). There were no statistically significant differences between the human-initiated

fluent the robot was during the collaborative tasks, the less users gazed to its face. We suggest that when the robot was acting more, the users focused on their logging task and their own actions rather than monitoring the robot.

Finally, we correlated gaze patterns and the objective metrics presented in Figure 10. We used again the Pearson product-moment correlation coefficient. We found that the average "robot only" time is correlated with the amount of gaze to the face (Cond H - Task A: $P_{coef} = 0.51$, $p < .05$; Task B: $P_{coef} = 0.53$, $p < .05$; Cond R - Task A: $P_{coef} = 0.53$, $p < .05$; Task B: $P_{coef} = 0.57$, $p < .05$; Cond P - Task A: $P_{coef} = 0.38$, $p > .05$; Task B: $P_{coef} = 0.23$, $p > .05$) and to the arms (Cond H - Task A: $P_{coef} = 0.47$, $p < .05$; Task B: $P_{coef} = 0.77$, $p < .05$; Cond R - Task A: $P_{coef} = 0.82$, $p < .05$; Task B: $P_{coef} = 0.60$, $p < .05$; Cond P - Task A: $P_{coef} = 0.66$, $p < .05$; Task B: $P_{coef} = 0.31$, $p > .05$). Next, we found correlation between "no motion" and gazes to face (Cond H - Task A: $P_{coef} = 0.48$, $p < .05$; Task B: $P_{coef} = 0.30$, $p > .05$; Cond R - Task A: $P_{coef} = 0.62$, $p < .05$; Task B: $P_{coef} = 0.46$, $p > .05$; Cond P - Task A: $P_{coef} = 0.67$, $p < .05$; Task B: $P_{coef} = 0.39$, $p > .05$). "Human idle" is also correlated to the amount of gazes to the robot's face (Cond H - Task A: $P_{coef} = 0.47$, $p < .05$; Task B: $P_{coef} = 0.53$, $p > .05$; Cond R - Task A: $P_{coef} = 0.44$, $p > .05$; Task B: $P_{coef} = 0.86$, $p < .05$; Cond P - Task A: $P_{coef} = 0.68$, $p < .05$; Task B: $P_{coef} = 0.19$, $p > .05$). As well as "robot idle" and gazes to face for Task A (Cond H - Task A: $P_{coef} = 0.48$, $p < .05$; Cond R - Task A: $P_{coef} = 0.63$, $p < .05$; Cond P - Task A:

help (H) and proactive robot (P) conditions in any of the sub-scales, despite the differences observed in objective metrics (e.g. the task completion time shown in Figure 9(b)) between these two conditions.

Subjective ratings of the quality of interaction appeared to be correlated with the number of actions performed by the robot (Figure 9(a)), rather than the overall task efficiency (Figure 9(b)). The reactive robot (R) condition was rated significantly lower than the other two (H and P) conditions, indicating that participants agreed significantly more that the quality was better in the H and P conditions (see Figure 13(a)). Whereas the significant differences were observed in the quality of interaction, participants did not rate differently the system performance. It seems they did not attribute the robot's behavior in the R condition to its inability to perceive the human or keep track of task progress.

In the forced ranking question administered at the very end of the study, 72% of participants (13/18) indicated P as their *most* preferred behavior, while 22% (4/18) indicated H and only 6% (1/18) indicated R. 78% of participants (14/18) indicated R as their *least* preferred behavior, with 17% (3/18) for H and 6% (1/18) for P. The question yielded a clear ranking of the three conditions as P > H > R from most preferred to least preferred. Furthermore, in separate two-choice questions, 67% of participants (12/18) indicated they prefer letting the robot take initiative, while the remaining 33% said they preferred having control over the robot's actions.

Fig. 13. Mean Likert-scale ratings in questionnaire responses. Significant differences according to Wilcoxon signed rank tests are

indicated with p-value ranges.

These results demonstrate that although there were no significant differences between the H and P conditions in the Likert-scale ratings, people are more likely to prefer P over H in favor of the improved objective metrics (Section 5.1).

5.3. Perceived differences of robot strategies

An open-ended question asked participants to describe the differences between the two conditions R and P in which the robot decided when to act, if they noticed any difference at all. All participants reported that they noticed a difference. The reactive robot was perceived as "slow" and characterized as "lazy" and "hesitant" by some of the participants. The proactive robot, on the other hand, was perceived "fast" and "pro-active". Descriptions of the perceived robot behaviors were accurate; for example:

- M, 35: "*... [P] felt more natural to have unprompted collaboration while I was performing the task, rather than the robot waiting for me to finish as it did during [R]*";
- M, 20: "*[P] was more **proactive** in its help ... [R], by contrast, would only complete actions that I was unable to complete*";
- M, 22: "*[In P] the robot took the **initiative** a lot more than [R]*".

These answers clearly highlight that the participants understood how the robot was behaving in the P condition, but did not feel that the robot was very motivated to help in the R condition. This understanding of the proactive robot's "mind" may be the reason why participants rated this condition as their favorite.

5.4. Collaboration enhancing human behaviors

The differences in the objective and subjective task metrics can be further dissected by examining the occurrence of certain events. Firstly, we saw that concurrent motion was significantly higher in the P and H conditions for Task A (Figure 10(c)), which shows better team work took place in these conditions. Secondly, in Task B, two objects (a container and a ball) were placed in the human-only zone. We observed that most people intuitively encouraged collaboration by starting tasks with objects that were in the human-only region of the table. Indeed, in the H and P condition, only three participants on average did not start with the ball in the human-only zone. In the R condition, seven users started with one of the balls placed in the both-allowed zone, showing lower collaboration. Participant descriptions of their strategies, in a free form question in the questionnaire, reflected their intent to enhance the collaboration; for example:

1. F, 22: "*[In R] I chose objects closest to me or that were obscuring the place of the objects needed to be. I also moved slower than I would without the robot to give it time to help*".
2. M, 19: "*[In H] I first wanted to set up the two bowls on the table before putting any of the balls in. This was to ensure that [the robot] would not attempt to put a ball in a space without a bowl*".
3. F. 24: "*[In P] I Moved the bowls and objects from the blue zone first and then help fill them [the bowls] one at a time*".

One participant placed the objects from his zone into the common central zone to make the robot perform the

task actions while he would perform the logging task. He described his strategy as:

1. M, 19: "*[In P] I moved objects from the blue zone into the collaboration zone, and placed objects in-between logging and [the robot's] actions*".
3. M, 35: "*I would ideally be able to give a couple of different command requests. The first command would be to move the pieces I cannot reach. The second command would be to just generally help out*".

6. Discussion

Our study demonstrated that the behavior of the proactive robot was similar to the behavior people asserted when they had control over the robot's actions. In turn, the similar high subjective rating of the proactive and the human-controlled robots could be partially ascribed to this similarity. Furthermore, our novel gaze analysis, we showed that the proactive and the reactive robots received less gazes to the face than the human-initiated robot (significantly in task B). It is suggested by psychologists that face-gaze (or eye-gaze) is an important component of social interactions (Argyle and Cook, 1976; Emery, 2000) and mutual understanding (Myowa-Yamakoshi et al., 2012; Tomasello et al., 2005), which contribute to natural interactions. In addition, we observed that the proportion of gazes to the face prior to a robot's action gets higher as the robot becomes less

thermore, we showed from the subjective ratings that the behavior that was common in these two conditions is more natural and fluent. Participants indeed described the robot's behavior during the proactive condition much better than for the reactive one, attesting of a better understanding of the robot's "mind".

On the question of *whether* a robot should take initiative, our results demonstrate that the answer depends on the robot's behavior. People would willingly give away control if the robot is proactive, but they would rather have control if it is reactive. Given its other benefits in terms of objective task and team metrics, this suggests that collaborative robots should be designed to always be proactive. Another interpretation is that users prefer robots that are more consistent and reliable, which was the case for the proactive robot. In this case, the predictability of the robot's actions should be improved to maximize the subjective and objective ratings.

In practice however, a proactive robot might not always be possible. Challenges such as partial task knowledge and uncertain perception might reduce the robot's ability to help the user when it is actually possible for it to help. While the simplistic help request used in our experiments would not be sufficient, enabling users to ask for particular types of help by commanding actions could result in more effective collaboration in such circumstances. This argumentation is supported by the answer to an open question asked during the final questionnaire: "How would you combine the different robot behaviors in an ideal robot assistant?". Some users answered that they would like a robot that first helps when asked or when needed, and can become more proactive on command.

1. F, 19: "*In an ideal robot assistant scenario, the robot would automatically decide to help when it was clear that the robot's help was needed; however, I would also be able to tell the robot when to help further*".
2. M, 20: "*I'd like to control the robot by giving him commands. Also he should react fast upon my calls. Maybe it would be better if I could tell him how to help me, like telling him which object to should he move, in this task for instance*".

autonomous, which appears to be related to turn taking cues as shown by Mutlu et al. (2009). Our correlation analysis results further revealed relations between gazes, intention understanding and turn taking.

Based on our gaze analysis results, we argue that if the robot is capable of taking turn autonomously, gazing to its face to give it turn (or to ask for help) would no longer be critical for efficient and natural interactions (e.g. lesser gazes to the face observed for the proactive robot). However, we suggest that gaze cues for turn taking and intention recognition should be used in scenarios where a robot is expected to interact *socially and* naturally with humans, while still being efficient and proactive.

The overall implication of our study is that mixed-initiative help triggers seems to be ideal for efficient collaborations in realistic settings. Increasing pro-activity over time, after observing the user's collaboration preferences (e.g. the work by Nikolaidis and Shah (2013)) might improve the collaboration, while benefit from the social aspect of the early human-initiated behavior. Finally, our study showed that human gaze toward the robot's face or arms can be interpreted as an intention cue or turn-taking signals that can be used to further improve the efficiency of human-robot interactions and lead to more natural collaboration.

From these various insights, we suggest two future work directions that could improve the efficiency of human-robot joint task collaboration: First, varying the robot's initiative based on experience and preference during the interaction with humans; second, taking into account gazes to the robot as turn-taking and intention cues to either trigger initiative or act upon specific objects desired by the human collaborator.

7. Conclusion

We developed a joint task execution system that autonomously performs a number of object manipulation tasks as well as monitoring end-to-end human task executions. Our system uses a dynamic Bayesian network to predict future environmental state and is capable of easily switching between various assistive behaviors.

We addressed the questions of whether and when a robot should take *initiative* during joint human-robot task execution by comparing three initiative models to trigger robot actions: *Human-initiated help*, *robot-initiated reactive help* and *robot-initiated proactive help*. Through a user study (N=18) we demonstrated that people collaborate best with a proactive robot, yielding better team fluency and high subjective ratings. While they are willing to give control of

Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, Proceeding of the Workshop on Cognition: A Bridge Between Robotics and Interaction. pp. 11. Portland, USA.

Cakmak M, Chao C and Thomaz A (2010) Designing interactions for robot active learners. *IEEE Transactions on Autonomous Mental Development* 2(2): 108–118.

Carifio J and Perla RJ (2007) Ten common misunderstandings, misconceptions, persistent myths and urban legends about Lik-

initiative to a proactive robot, they prefer having control rather than working with a reactive robot that only helps when it is needed.

Additional evidences showed that participants gazed to the robot's face more often during the *human-initiated help* than for the other conditions. We also showed that participants almost always gazed to the face of the robot before asking for help, which can be used as a cue for turn taking and improve the robot's reaction time. This may mean that asking for the robot's help may lead to a more "social" interaction, without altering the quality of interaction or the system performance.

Acknowledgements

The authors would like to thank Dan Butler for helping in the implementation of the perceptual system.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by MEXT/JSPS Grants-in-Aid for Scientific Research (grant numbers 24119003, 24000012 and 25700027), JSPS Core-to-Core Program, the ONR Science of Autonomy (grant number N000141310817) and NSF (grant number 1318733).

References

Alexandrova S, Cakmak M, Hsiao K, et al. (2014) Robot programming by demonstration with interactive action visualizations. In: *Proceedings of Robotics: Science and Systems*. Berkeley, CA.

Argyle M and Cook M (1976) *Gaze and Mutual Gaze*. American Psychological Association. Cambridge University Press.

Awais M and Henrich D (2012) Proactive premature intention estimation for intuitive human-robot collaboration. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4098–4103. Vilamoura-Algarve, Portugal: IEEE.

Baraglia J, Cakmak M, Nagai Y, Rao, R and Asada M (2016) Initiative in robot assistance during collaborative task execution. In: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 67–74. Christchurch, New Zealand: IEEE.

Baraglia J, Nagai Y and Asada M (2014) Prediction error minimization for emergence of altruistic behavior. In: *The 2014 4th International Conference on Development and Learning and on Epigenetic Robotics*, pp. 281–286. Genoa, Italy: IEEE.

Baraglia J, Nagai Y and Asada M (2015) State prediction for development of helping behavior in robots. In: *The 2015*

ert scales and Likert response formats and their antidotes. *Journal of Social Sciences* 3(3): 106–116.

Chao C and Thomaz A (2013) Controlling social dynamics with a parametrized model of floor regulation. *Journal of Human Robot Interaction* 2(1): 4–29.

Chao C and Thomaz AL (2010) Turn taking for human-robot interaction. In: *AAAI fall symposium: Dialog with robots, The 2010 AAAI Fall Symposium*, USA. Arlington, USA: AAAI press.

St Clair A and Mataric M (2015) How robot verbal feedback can improve team performance in human-robot task collaborations. In: *Proceedings of the 2015 Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 213–220. Portland, USA: ACM.

Cuntoor NP, Collins R and Hoogs AJ (2012) Human-robot teamwork using activity recognition and human instruction. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.459–465. Vilamoura-Algarve, Portugal: IEEE.

Degris T and Sigaud O (2010) Factored Markov decision processes. In: *Markov Decision Processes in Artificial Intelligence*. pp.99–126. Elsevier, Open access.

Dragan AD, Bauman S, Forlizzi J and Srinivasa SS (2015) Effects of robot motion on human-robot collaboration. In: *Proceedings of the 2015 Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pp.51–58. Portland, USA: ACM.

Emery NJ (2000) The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews* 24(6): 581–604.

Fong T, Thorpe C and Baur C (2003) Multi-robot remote driving with collaborative control. *IEEE Transactions on Industrial Electronics* 50(4): 699–704. IEEE.

Gombolay MC, Gutierrez RA, Sturla GF, et al. (2014) Decision-making authority, team efficiency and human worker satisfaction in mixed human-robot teams. *Robots: Science and Systems (RSS)*.

Groten R, Feth D, Peer A and Buss M (2010).Shared decision making in a collaborative task with reciprocal haptic feedback-an efficiency-analysis. In: *Robotics and Automation (ICRA), IEEE International Conference on 2010*, pp. 1834–1839. Anchorage, USA: IEEE.

Hawkins KP, Bansal S, Vo NN and Bobick AF (2014) Anticipating human actions for collaboration in the presence of task and sensor uncertainty. In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp.2215–2222). Hong Kong, China: IEEE.

Hayes B and Scassellati B (2015) Effective robot teammate behaviors for supporting sequential manipulation tasks. In: *Intelligent Robots and Systems (IROS), IEEE/RSJ International Conference on 2015*, pp. 6374–6380. Tokyo, Japan: IEEE.

Hoffman G (2013) Evaluating fluency in human-robot collaboration. In: *HRI workshop on human robot collaboration*.

Hoffman G and Breazeal C (2007, March). Effects of anticipatory action on human-robot teamwork efficiency, fluency, and per-

Mutlu B, Terrell A and Huang CM (2013) Coordination mechanisms in human-robot collaboration. In: *Proceedings of the*

ception of team. In: *Proceedings of the 2007 2nd ACM/IEEE international conference on Human-robot interaction*, pp.1–8. Washington DC, USA: ACM.

Hoffman G and Breazeal C (2010) Effects of anticipatory perceptual simulation on practiced human-robot tasks. *Autonomous Robots* 28(4): 403–423.

Huang CM and Mutlu B (2014) Learning-based modeling of multimodal behaviors for humanlike robots. In: *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pp.57–64. Bielefeld, Germany: ACM.

Huang CM, Andrist S, Sauppé A, et al. (2015a) Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology* 6: 1049.

Huang CM, Cakmak M and Mutlu B (2015b) Adaptive coordination strategies for human-robot handovers. In: *2015 Robotics: Science and systems (RSS)*. Ann Arbor, Michigan, USA.

Jarrassé N, Paik J, Pasqui V and Morel G (2008) How can human motion prediction increase transparency? In: *Robotics and Automation, ICRA 2008. IEEE International Conference on 2008*, pp.2134–2139. Pasadena, USA: IEEE.

Kwon WY and Suh IH (2013) Proactive planning using a hybrid temporal influence diagram for human assistive robots. In: *Robotics and Automation (ICRA), IEEE International Conference on 2013*, pp.1785–1791. Karlsruhe, Germany: IEEE.

Li H, Cabibihan JJ and Tan YK (2011) Towards an effective design of social robots. *International Journal of Social Robotics* 3(4): 333–335.

Mainprice J and Berenson D (2013) Human-robot collaborative manipulation planning using early prediction of human motion. In: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.299–306. Tokyo, Japan: IEEE.

Mainprice J, Sisbot EA, Jaillet L, Cortés J, Alami R and Siméon T (2011) Planning human-aware motions using a sampling-based costmap planner. In: *Robotics and Automation (ICRA), IEEE International Conference on 2011*, pp.5012–5017. Shanghai, China: IEEE.

Meltzoff AN (2007) 'Like me': A foundation for social cognition. *Developmental Science* 10(1): 126–134.

Moon A, Troniak DM, Gleeson B, et al. (2014). Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing. In: *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pp.334–341. Bielefeld, Germany: ACM.

Mutlu B, Shiwa T, Kanda T, Ishiguro H and Hagita N (2009) Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In: *Proceedings of the 2009 4th ACM/IEEE international conference on Human robot interaction*, pp.61–68. San Diego, USA: ACM.

2013 Workshop on Collaborative Manipulation, 8th ACM/IEEE International Conference on Human-Robot Interaction. Tokyo, Japan: ACM.

Myowa-Yamakoshi M, Scola C and Hirata S (2012) Humans and chimpanzees attend differently to goal-directed actions. *Nature Communications* 3: 693.

Najmaei N and Kermani MR (2010) Prediction-based reactive control strategy for human-robot interactions. In: *2010 IEEE international conference on robotics and automation (ICRA)*, pp.3434–3439. Anchorage, USA: IEEE.

Nikolaidis S, Ramakrishnan R, Gu K and Shah J (2015) Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. In: *Proceedings of the 2015 Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pp.189–196. Portland, USA: ACM.

Nikolaidis S and Shah J (2013) Human-robot cross-training: Computational formulation, modeling and evaluation of a human team training strategy. In: *Proceedings of the 2013 8th ACM/IEEE international conference on Human-robot interaction*, pp.33–40. Tokyo, Japan: IEEE Press.

Pérez-D'Arpino C and Shah JA (2015) Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classification. In: *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp.6175–6182. Seattle, USA: IEEE.

Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classification. In: *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp.6175–6182. Seattle, USA: IEEE.

Shah J, Wiken J, Williams B and Breazeal C (2011) Improved human-robot team performance using chaski, a human-inspired plan execution system. In: *Proceedings of the 6th international conference on Human-robot interaction*, pp.29–36. Lausanne, Switzerland: ACM.

Sisbot EA, Clodic A, Alami R and Ransan M (2008) Supervision and motion planning for a mobile manipulator interacting with humans. In: *Human-Robot Interaction (HRI), 3rd ACM/IEEE International Conference on 2008*, pp.327–334. Amsterdam, Netherlands: IEEE.

Tomasello M, Carpenter M, Call J, et al. (2005) Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences* 28(05): 675–691.