A Framework to Explore Proximate Human-Robot Coordination

SACHIKO MATSUMOTO, AURIEL WASHBURN, and LAUREL D. RIEK*, Computer Science and Engineering, UC San Diego

Proximate human-robot teaming (pxHRT) is a complex subspace within human-robot interaction. Studies in this space involve a range of equipment and methods, including the ability to sense people and robots precisely. Research in this area draws from a wide variety of other fields, from human-human interaction to control theory, making study design complex, particularly for those outside the field of HRI. In this paper, we introduce a framework that helps researchers consider tradeoffs across various task contexts, platforms, sensors, and analysis methods; metrics frequently used in the field; and common challenges researchers may face. We demonstrate the use of the framework via a case study which employs an autonomous mobile manipulator continuously engaging in shared workspace, handover, and co-manipulation tasks with people, and explores the effect of cognitive workload on pxHRT dynamics. We also demonstrate the utility of the framework in a case study with two groups of researchers new to pxHRT. With this framework, we hope to enable researchers, especially those outside HRI, to more thoroughly consider these complex components within their studies, more easily design experiments, and more fully explore research questions within the space of pxHRT.

CCS Concepts: • **Embedded and cyber-physical systems → Robotics**; • **Human-centered computing** → *Interaction design theory, concepts and paradigms*; • **Computing methodologies** → *Artificial intelligence.*

ACM Reference Format:
Sachiko Matsumoto, Auriel Washburn, and Laurel D. Riek. 2022. A Framework to Explore Proximate Human-Robot Coordination. 1, 1 (February 2022), 35 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Robots are increasingly moving from well-controlled spaces into human-centered environments, including hospitals, homes, and disaster zones. These spaces are more dynamic and complex than the laboratory and industrial settings robots have traditionally been used in and are often safety critical.

Therefore, in order to perform well and decrease the risk to people in these environments, good teaming between people and robots is essential. Such teaming will enable robots and humans to augment each others' abilities to perform more complex functions.

In order to develop methods to allow robots to do this, both researchers and robots need methods to characterize interactions with human partners. While the exact elements that constitute effective teaming are context-dependent, if researchers have techniques to well-characterized aspects of the interaction, it will enable them to better understand how robots could react more appropriately and fluently in a given situation. In turn, adaptations of these techniques could be used in robot control architectures to improve their interactions with people. Without methods to characterize

Authors' address: Sachiko Matsumoto; Auriel Washburn; Laurel D. Riek, smatsumo@eng.ucsd.edu, Computer Science and Engineering, UC San Diego, La Jolla, CA.

XXXX-XXXX/2022/2-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

and evaluate the effects of their behavior, robots may inadvertently act in ways that degrade the quality of their interactions in human-robot teams.

Dynamical and complex systems analysis techniques have emerged as powerful tools to assess coordination in a variety of settings. Researchers have applied some of these techniques in HRI research, such as Granger Causality, Cross-Recurrence Quantification Analysis, and cross correlation. [30, 38, 41].

The majority of work in proximate human-robot teaming (pxHRT) has focused on shared workspace tasks [13, 21, 24, 53, 73] and handovers [11, 69], with relatively little on co-manipulation tasks [17, 51]. Furthermore, in HRT, standard metrics have only recently been studied [23].

While there is a large body of work in proximate, physical HRT, to our knowledge there is no systematic approach to design and run studies in this space. While some articles provide guidance for HRI research (e.g., [6, 26, 27, 55]), little guidance targets pxHRT specifically. For instance, frameworks exist for social HRI [3], user-centered HRI [32], and levels of robot autonomy [5], among others. Onnasch and Roesler [55] recently proposed a taxonomy to assist with analyzing HRI studies. Their framework includes three clusters: interaction context, robot characteristics, and team classification. However, their taxonomy is more suited for retrospective applications, e.g., enabling researchers to examine prior studies to better compare and contrast HRI work. In contrast, the framework we introduce in this paper is designed for prospective use, and specifically targeted to pxHRT research.

Some recent work has also provided guidance specifically on designing HRI studies. For example, Hoffman and Zhao [26] provide a thorough guide to designing HRI studies focused on experimental research, from deciding on a research question, to designing the study, to performing statistical tests and reporting results. Belpaeme [6] similarly offers advice to new HRI researchers, discussing some critical decisions that need to be made during experimental design, as well as common mistakes in data analysis. Both of these papers are extremely useful guides to new researchers designing experiments. However, they discuss HRI research as a whole and do not touch on issues that are of particular importance to pxHRT study design. We hope that our framework will complement these other guides by providing advice explicitly aimed at pxHRT researchers.

In our framework, we suggest several steps and tradeoffs to consider while designing a pxHRT study, and provide examples from the literature to illustrate how others have dealt with these tradeoffs. We also discuss common challenges that occur specifically in pxHRT studies. We understand that many senior HRI researchers will already be familiar with the options and tradeoffs that our framework introduces. However, we note that HRI is a relatively small subfield in robotics, and many roboticists who do not specialize in HRI still conduct HRI studies, particularly pxHRT research. This framework is geared toward that audience.

We demonstrate how our framework can be used by describing a case study from our own research. We designed the study to inform the metrology of teaming by investigating the dynamics between a person and robot performing several pxHRT tasks under different workload conditions. We use this study to demonstrate how our proposed framework could be employed to design pxHRT studies.

We further demonstrate the potential utility of our framework in a case study with two groups of researchers new to the field of pxHRT. The researchers used our framework in the design of a study and provided feedback via a short, semi-structured interview.

The framework we present will enable researchers to better navigate the pxHRT space and design studies to analyze coordination in human-robot teams across different task contexts. The framework is general enough to be used to investigate a wide array of challenges in a variety of contexts. Running a study requires careful, thoughtful planning. Moreover, there are many design considerations in pxHRT studies that researchers new to the area may not be aware of. By providing

structure for these considerations, we hope this framework will ease the study design process, enabling researchers to better use the time spent designing their studies, and thus produce more well-thought-out studies. This, in turn, will potentially reduce the chance for unnecessary errors along the way. Additionally, by providing researchers with a structure for study design, we hope this framework will assist researchers in designing easily reproducible, transparent studies. We hope that this will accelerate advancements in pxHRT research.

2 BACKGROUND

2.1 Study Design and Frameworks in HRI

Some recent papers have addressed study design in HRI broadly [6, 26]. Belpaeme [6] goes over important decisions during HRI study design, including level of autonomy, lab vs. non-lab studies, and single vs. long-term interactions. Belpaeme also discusses several problems that can occur in collecting and analyzing data, such as concerns about null-hypothesis significance testing, selective data publication, and the Hawthorne effect. Hoffman and Zhao [26] provide detailed steps for experimental HRI studies. They break the process down into 10 stages, from research questions to reporting, and provide suggestions and considerations for each stage.

As good study design principles for HRI clearly also apply to pxHRT, our framework naturally overlaps with these guidelines for study design. However, because we focus on pxHRT, our framework includes elements that are not explicitly discussed in these broader guidelines. Thus, this work complements these guidelines and provides additional support for researchers new to pxHRT.

Researchers have also proposed frameworks for HRI research generally. Onnasch and Roesler [55] created a taxonomy to classify studies based on interaction context, robot characteristics, and teams. This can be useful for designing studies, but also enables researchers to better compare and contrast previous studies. Yanco and Drury [85, 86] proposed a taxonomy for classifying human-robot interactions that included factors such as task type and criticality, level of shared interaction among teams, and decision support for operators. Parashar et al. [57] reviewed previous HRI taxonomies and used this to develop a taxonomy divided into system context, local dynamics, and effects.

Additionally, some papers present guidance or frameworks for specific areas of HRI research. These papers provide more details that are relevant to the areas they focus on, which may not be applicable to other areas of HRI. For instance, Baraka et al. [3] proposed a framework for classifying social robots. Similarly, Beer et al. [5] developed a framework for categorizing levels of robot autonomy, drawing on classifications of levels of autonomy from human-automation literature. Researchers also developed frameworks for user-centered HRI studies [32] and approaches for HRI experiments with humanoid robots [27]. Like these papers, we propose a framework to guide study design in a specific subfield of HRI, namely pxHRT.

In many fields, iterative study designs are common, where researchers explore their data then make changes to many aspects of their experimental design[65]. This practice is particularly common in context-rich, dynamic experimental settings, such as schools and health settings [31, 66]. For example, in one of our experiments on human-robot trust and setting expectations [76], we needed to determine exactly when in the study a robot should make an error. We iteratively collected small batches of pilot data, analyzed it, then made slight modifications to our study design. Once we saw effects more reliably, we finalized our study design and recruited participants for the main experiment. Therefore, to represent this important aspect of study design, we include an iterative loop around our framework.

Researchers also addressed practices specifically in the measurement and analysis of data. For instance, Schrum et al. [64] found that HRI researchers rarely use Likert scales completely properly

and developed recommendations for researchers to improve their use of such scales. Additionally, many researchers are concerned about current p-value usage [6, 15], as well as the possibility for researchers, intentionally or not, to get significant results, regardless of whether or not an effect exists [68]. To avoid some of these problems, they recommend fully describing an experimental design before collecting data and pre-registering the experiment (e.g., with the Open Science Foundation) [6, 15, 68]. Any deviations from the planned experiment during or after data collection should be explained in the paper.

2.2 Coordination in HRI

Coordination is a critical component of joint action, and thus will be important in many emerging opportunities for proximate human-robot teaming. However, designing robots to coordinate well with humans is difficult, particularly in human-centered spaces that are not explicitly designed to accommodate robots.

Furthermore, there is a large amount of variance in the human population, so different people may perform the same task in multiple ways. Rather than having a single strategy to coordinate with people on a given task, robots must adapt their technique to the specific people they work with. Researchers have investigated many strategies to individualize robotic behavior to a given partner [24, 48, 51, 53]. For example, Nikolaidis and Shah [53] proposed a cross-training technique in which a robot and person iteratively switch roles in a task so they learn how to best perform the task together.

To date, much work on joint action has focused on shared workspace tasks (e.g., [13, 21, 24, 53, 73]), as well as handovers [11, 69], with some work on co-manipulation (e.g., [17]). pxHRT research intersects many different fields, from human factors to design to control theory. This can make the literature space difficult to navigate, as many papers examining the same problem approach it from widely varying perspectives. For instance, Hoffman and Breazeal [25] and Nikolaidis and Shah [53] both explore how humans and robots can more efficiently and fluently coordinate with each other. However, Hoffman and Breazeal [25] focused on a cognitive architecture for the robot to enable it to better anticipate the person's intent. Nikolaidis and Shah [53], on the other hand, focused on a training method for the human and robot to allow them to have a better shared model of the task.

Conversely, researchers often use similar tasks to investigate vastly different questions. Peternel et al. [58] and Reed et al. [60] both employ a cooperative crank turning task. Peternel et al. [58] use this task to explore a control framework, whereas Reed et al. [60] characterize human-human teams and discover different dynamics between human-human and human-robot teams.

Thus, diversity and complexity of the field, while useful for designing new methods for teaming, can also make it harder to fully grasp. It can also hinder experimental design. Our framework attempts to address these issues by suggesting methods for researchers to traverse this space.

2.3 Measuring Movement Coordination in Human-Human Interaction

Our framework includes different types of coordination analyses, many of which originated in human-human interaction studies. For readers not familiar with these techniques, we provide background here.

Research aimed at understanding human-human interaction incorporates many different approaches to evaluating the coordination of inter-agent behaviors. Together these approaches provide a wide range of information about interpersonal interactions. For example, they can reveal differences in movement coordination associated with dancing expertise [77], as well as differences in the temporal coordination of musical behaviors based on a musician's role within an ensemble [83]. Coordination measures can also identify social cohesion and connection between human team members. Researchers in interpersonal interaction have established strong connections between

interpersonal movement coordination and reports of team-member liking and affiliation [43, 45] as well as social attachment and cooperation [82]. They also found that coordinated movement can have positive effects on collaborative problem-solving during teaming [44].

Measuring interpersonal coordination characteristics can provide reliable and meaningful information about human interaction. Higher levels of coordination are associated with better rapport between individuals [43, 45], as well as better cooperation and collaboration during teamwork [44]. Recent work also demonstrated that individuals observing human dyads expect that dyads who exhibit synchrony will work well together and that observers are also more interested in affiliating with these dyads [40]. Additionally, the range and frequency of phase relationships between repetitive movements conveys information about the stability of coordination over a given period of time. Specifically, for highly stable coordination a limited range of phase relationships are visited with high frequency, while for unstable coordination a wide range of phase relationships may be visited at relatively equal, often low, frequencies. For tasks that depend on the coordination of joint actions between individuals performance is typically better if coordination is stable.

Beyond the general magnitude and stability of coordination, researchers can also evaluate the influence interacting individuals have on each other's behavior. Interpersonal coordination depends on the fact that one individual has information about another's behavior, such that they are unidirectionally coupled to their actions. Commonly, all pairs of individuals engaged in an interaction have information about each other and are therefore bi-directionally coupled. Behavioral coordination arising from bi-directionally coupled individuals is characterized by mutual adaptation, in which both actors adjust their actions to account for those of their co-actor.

Coordination characteristics like magnitude, stability, mode, and adaptation can also change over the course of interpersonal interaction. As a result, the evolution of spatiotemporal coordination characteristics can be as informative as the characteristics themselves. For example, continual increases in the variability of relative phase relationships exhibited by two interacting individuals over time predict an eventual inability to maintain the current coordinated state. Schmidt et al. [63] observed this when they asked pairs of individuals to coordinate rhythmic leg movements in an anti-phase pattern and introduced increases in movement frequency, leading to steady increases in relative phase variability followed by a breakdown in the anti-phase coordination pattern.

In addition to the information one can gain from assessing the evolution of close spatiotemporal coordination, one can also understand more about an interaction from looking at shared behavioral structure between interacting individuals. By understanding the properties of human behavior as consistent with those of dynamical systems, it is possible to render its evolution in reconstructed phase space.

With this perspective, researchers developed cross-recurrence quantification analysis (CRQA) to identify the presence and duration of overlaps in the behavioral dynamics of concurrent behaviors displayed by interacting individuals, quantifying the regularity, stability, predictability, and homogeneity of an interaction (see [67, 78, 79] for detailed descriptions). Granger Causality (GC) is another commonly used technique to examine relationships in time series data. It indicates when information from one series is dependent on information from another series (see [4] for a detailed explanation and software tools). For instance, it can be used to determine leader-follower relationships [12].

Understanding movement coordination patterns contributes significantly to our ability to interpret many aspects of interpersonal interactions. The same approach can be used within human-robot teaming. A few groups have already centered the importance of movement coordination within human-robot interaction research to: propose the development of greater movement adaptation abilities for social robots bidirectionally coupled to human co-actors [37–39], use oscillatory behavioral dynamics to control robot behavior during proximal human-robot interaction [35, 49],

measure human movement behavior via mobile robot sensors [30], and use information about human group coordination to plan and perform cohesive robot movements [29].

3 FRAMEWORK

Physical pxHRT is a large space with many different research questions, platforms, sensors, study designs, measures, and analysis methods to choose from. Across each of these, researchers must consider a wide variety of tradeoffs in terms of cost, time, and complexity. Additionally, during HRT studies, numerous other challenges manifest themselves, including problems involving hardware, systems, people, and data labelling and analysis. This makes it difficult to design and run studies while balancing these tradeoffs and challenges.

In this paper, we introduce a new conceptual framework to help pxHRT researchers navigate this space and support them in their efforts to design new lines of work. We derived this framework from insights from the literature, as well as from our own experimentation. We suggest steps for designing a study, and provide examples from the literature to elucidate each step.

Our framework consists of seven components (Fig. 2): the research question, task context, platform, sensors, autonomy, evaluation, and analysis methods. Below, we describe the primary connections between these components, as depicted by the grey arrows in Fig. 1. We emphasize that all of the components of the study are interconnected to some extent, so any given component should not be thought of as completely isolated from any other. However, we believe that the connections described here are the ones with the most influence and should be carefully considered by researchers during study design.

The *research questions* are particularly important as they will drive many other decisions during the study design. This refers to the question the researcher is trying to answer by running the study. For instance, a research question about whether spatial and temporal contrast affect human-robot handover fluency will necessitate a study design that includes a handover [11]. Similarly, a research question about how well a new algorithm enables a robot to team with groups of people will result in an evaluative paradigm in which the robot interacts with at least two people at a time [71, 74]. These questions are typically human-focused, robot-focused, or interaction-focused. For example, a question about how humans and robots collaborate physically via haptic communication is interaction-focused [60], while a question about how fluent people perceive a human-only vs. a human-robot team to be is human-focused [21]. A question about how well a data-efficient reinforcement learning approach can model physical human-robot interactions is robot-focused [20].

The *task context* researchers choose will also affect several components, especially the platform and sensors used for the study, as well as which analysis methods are most appropriate. The task context includes the type of task used in the study, as well as how it is represented, particularly whether it is continuous or discrete. Possible types of tasks include handovers (passing an object between agents), co-manipulation (two or more agents moving the same object at the same time), and share workspace tasks (two or more agents performing a task in the same space). If the researchers want a co-manipulation task context, where two agents manipulate an object at the same time, they will need a platform with a manipulator, whereas if they want a shared workspace task context, a non-manipulator mobile robot or even a tabletop platform may be sufficient. Additionally, the task context, in conjunction with the research question, will affect what types of sensors a researcher needs. For example, in a handover task, a researcher may be interested in carefully tracking a person's hand movements, while for a shared workspace task, they may want to know the position of the person's body but not be concerned about the person's hands specifically.

While some analysis methods can be applied across almost any task context, others are narrower in scope. An analysis of idle time may make sense for a shared workspace task, but is unlikely to matter in a co-manipulation task, where the person and robot are both continuously active.

The *platform* and *sensors* that researchers choose to use will often be the resources already available to them. For instance, researchers may already have certain robots or sensors in their lab, in which case they may preferentially design experiments using these hardware components. For example, they may have access to a platform with a mobile base but no manipulators, which might push them towards a shared workspace task rather than a co-manipulation task. Thus, the platform researchers decide to use will affect the task context.

Which sensors researchers choose may affect the level of robot autonomy and the analyses researchers can perform. For instance, if the robot does not have access to sensor data that can accurately determine the location of a person's hand, it may be difficult for the robot to be fully autonomous during a handover task. Additionally, some data processing, such as modeling human arm motions, may require specific types of data, such as sensor measurements from different points on a person's arm [58]. Similarly, if a researcher wants to run analyses on the proxemics of the person and robot, they will need to use sensors that will allow them to know the positions of both.

The level of *autonomy* of the robot is how much the robot does on its own versus how much it is controlled by an operator. The autonomy will affect and be affected by the sensors the researcher chooses, and may also impact the evaluative paradigm. For instance, to learn and imitate force constraints for a collaborative task, the robot must have force sensors [62]. Similarly, if a robot needs to autonomously coordinate with people during a synchronous motion task, like dancing, it might need a way to visually sense those people [14, 29]. Furthermore, the task context will affect autonomy, as a more complex task, such as performing surgery, may be more difficult or dangerous to automate.

The level of robot autonomy may also affect the *evaluation*. The evaluation includes the details of how the research question will be evaluated, such as the specific conditions that will be used, the population participants will be drawn from, the specifics of how the person and robot interact, and so on. The degree of autonomy could exclude some interactions that are beyond current limitations of robots. For instance, a robot may have difficulty autonomously picking up a novel object from an extremely cluttered environment. The evaluation paradigm also impacts the analysis methods that researchers can use. If a researcher chooses to have relatively short trials, they may not be able to use analyses like cross-correlation that require longer trial lengths. Thus, researchers should consider what evaluative paradigms best allow them to answer their research question, while also taking the desired levels of autonomy and analysis methods into account.

Finally, the *analysis methods* consist of the data being collected and the types of techniques used to analyze that data. They are most strongly impacted by the sensors, evaluation, and task context, as discussed previously. When designing a study, we hope that taking these connections into account will assist researchers in determining the best design for their study.

In the following subsections, we will describe each component of our framework in more detail, as well as important categories within each component, as listed in Fig. 2.

3.1 Research Questions

The first consideration of the framework are the kinds of questions an HRI researcher might ask within the context of teaming. There are many ways researchers might generate a research question. Often, researchers review relevant literature in the field and identify gaps where there are still open questions. They may also devise a research question based on their own prior work. Similar to our delineations within the HRI conference, these questions may be more technically-focused, human-focused, and/or interaction-focused in scope. Technically-focused questions often attempt

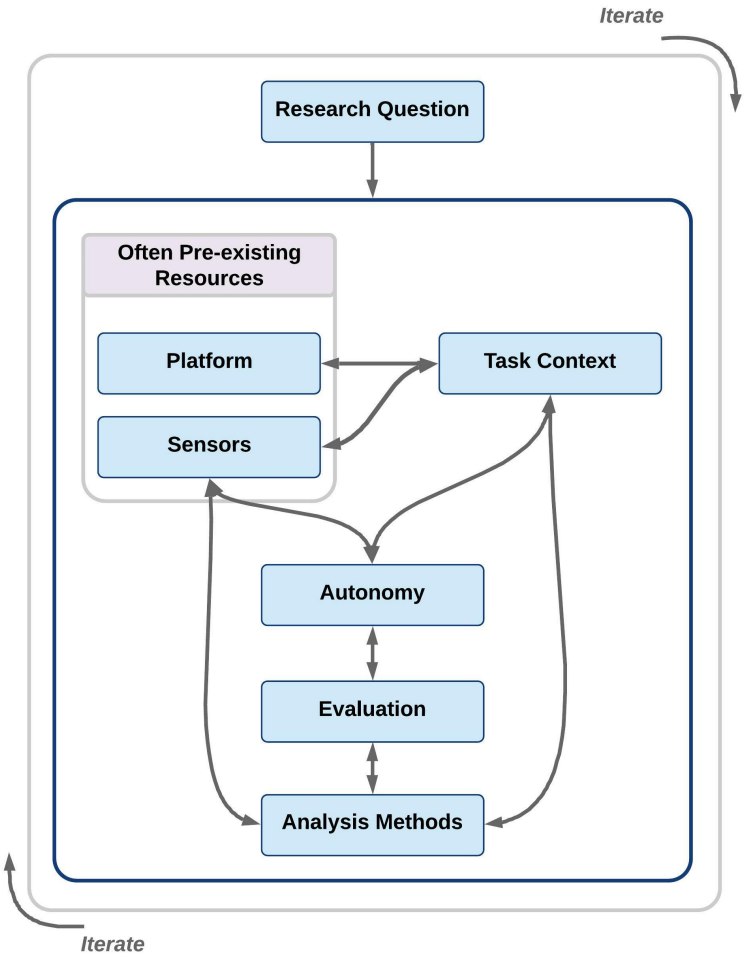

Fig. 1. Visualization of our framework for guiding researchers in the design of pxHRT studies. Grey arrows show the main connections between components, though no component is completely independent from any other component. Additionally, researchers will likely iterate on their study design several times.

to improve the capabilities of robots. pxHRT technical questions often involve designing a control paradigm for a robot and comparing it against a baseline. There are no standard baselines for many pxHRT tasks, so researchers will typically compare against another method that is commonly used for the task. Other times, they will compare against a robot that has a fixed policy [48].

Many researchers have explored technically-focused questions [16, 17, 25, 29, 41, 46, 48, 52, 53, 56, 70]. Some choose to look at higher level architectures, such as the method the robot uses to learn about and encode the task or world state. For example, Nikolaidis et al. [53] developed a cross-training method to enable humans and robots to converge to a shared model of a collaborative task. They then evaluated their model against a reinforcement learning algorithm in which the human manually assigned rewards to robot actions. Other researchers focus on low-level controllers to enable robots to better engage in pxHRT. Peternel et al. [58] developed a hybrid controller that

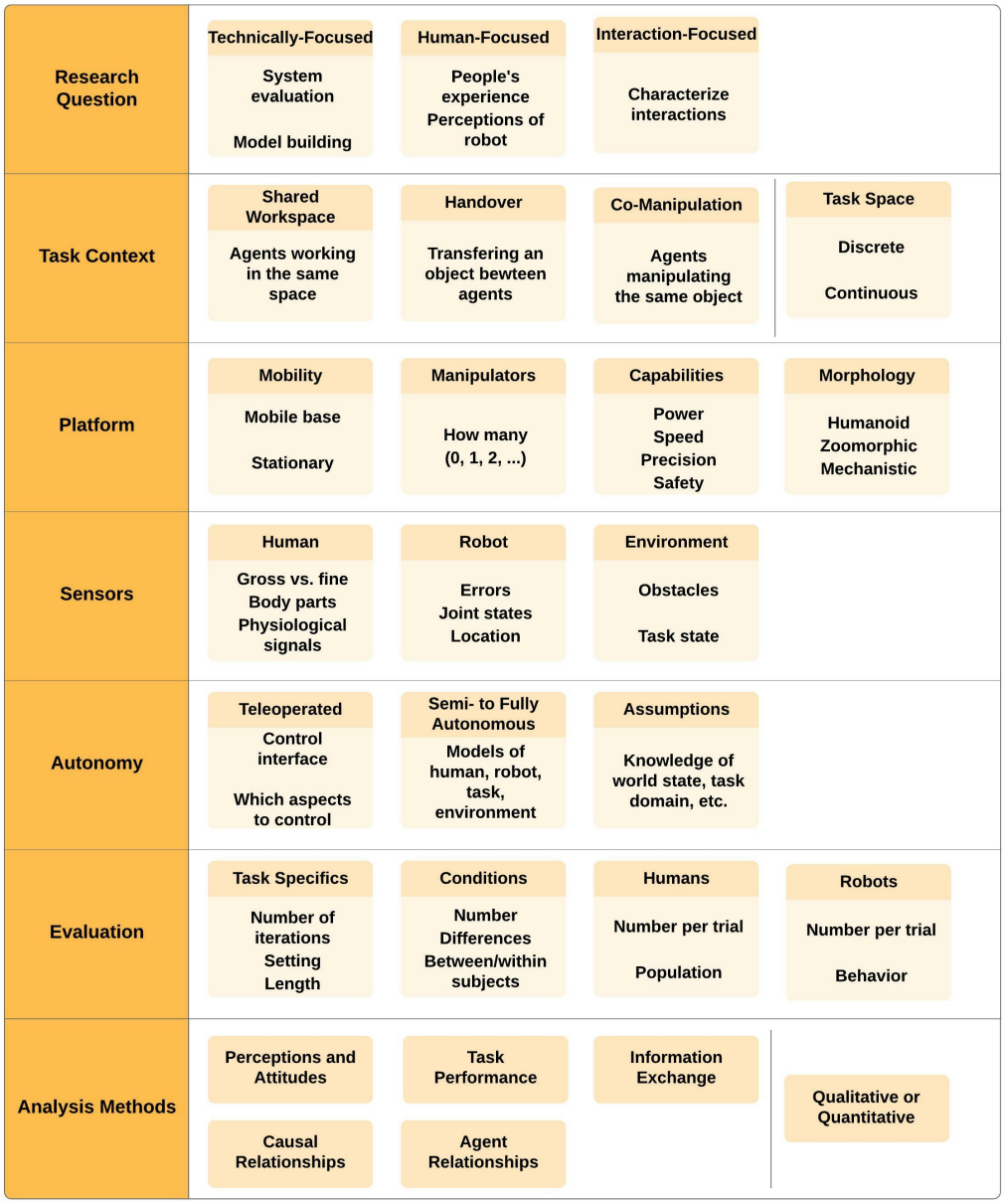

Fig. 2. Important categories within each component in our proposed framework.

took the person's arm stiffness into account. Technical questions are valuable for improving robot function and extending robot capabilities. However, they often need to be informed by research on human-focused and interaction-focused questions in order for the robot to act in a way people find acceptable.

Human-focused questions generally investigate people's experiences with and perceptions of robots. While these studies do not usually directly generate new robot capabilities, they can provide insight into what capabilities robots need to collaborate with humans, and how current capabilities should be integrated to improve pxHRT. In these cases, researchers might design an experiment in which people interact with other people, robots, or both and examine the effects of different conditions on people's experience. For example, researchers have studied people's reactions to robot versus human assistants in industrial settings [21, 73], as well as when performing a co-manipulation task with a human [80]. Gombolay et al. [21] investigated the effects of decision making authority in human teams compared to human-robot teams during a shared workspace task. They then provided recommendations for deploying robots into human work environments, based on their results. These recommendations could be used as guidelines for desired robot behavior in the design of controllers for pxHRT.

Finally, some researchers explore interaction-focused questions in their pxHRT research [35, 38, 41, 60]. Interaction-focused questions typically focus on characterizing the interactions among human and/or robot team members. Results from these studies can provide insights into metrics that can be used to evaluate and characterize interactions, as well as behaviors robots should exhibit during proximal, physical teaming, which can inform controller design. For instance, Reed et al. [60] examined the forces people display during a cooperative task when they can only communicate haptically. They found that the force profiles were different when a person was paired with another person than when they interacted with a robot that imitated their force profile, even when they believed the robot to be another human. They also were able to characterize human dyads based on the way they specialized their force output, which could be used to design more human-like controllers for human-robot haptic interactions. Medina et al. [41] collected data from human-human handovers to design a controller for human-robot handovers. They used insights from the human-human handovers to predict the desired trajectory and handover pose during human-robot handovers.

The type of question a pxHRT researcher chooses to investigate will have a major impact on the rest of their study design. It may affect whether only humans are involved, as could be the case in an interaction-focused study, or if a robot is required, as is generally the case in a technically-focused study. It also will determine whether different modes of robot behavior are needed, as would often be the case in a technically-focused study, but not necessarily in an interaction-focused study. Furthermore, when choosing their research question(s), researchers may want to consider their level of expertise in different areas of pxHRT. For instance, if pursuing a technically-focused question, experience with collaborative control architectures could be useful. On the other hand, knowledge of time series and coordination analyses might be more relevant for an interaction-focused question. The question chosen will influence the way in which the study will contribute to proximate human-robot collaboration, whether by improving robot capabilities, providing insights into people's experience with robots, or elucidating new ways to characterize interactions. Thus, the type of research question(s) a pxHRT researcher asks will impact the rest of their decisions for the study.

3.2 Task Context

After deciding on a research question, the researcher will likely want to determine the task context of interest. This is an important decision in physical, proximate human-robot collaboration studies because different task contexts require different methods of interaction between team members. For example, constructing an object with someone and carrying a table with them are both forms of proximate teaming. However, the dynamics involved and types of sensing required are very

different for those two tasks. Hoffman [23] identifies three distinct task contexts which play a role here: shared workspace tasks, handovers, and co-manipulation tasks [11, 13, 21, 24, 29, 53, 69, 73].

Shared workspace tasks involve two or more agents working on a given task in the same space. They may perform different subtasks within the task or perform the same subtasks in parallel.

For instance, in Nikolaidis and Shah [53] the person placed screws, which the robot then drilled in. Additionally, Iqbal et al. [29] examined human-robot synchrony while dancing. A robot joined a group of humans performing a simple dance and attempted to complete the moves at the same time as the people.

Handovers are the direct transfer of an object between two agents, that is without setting the object down during the transfer. Within this space, people focus on different aspects of handovers. For instance, Cakmak et al. [11] explored how making the handover pose spatially distinct and temporally distinct from a carrying pose affected the timing of handovers. Moon et al. [47] investigated how gaze behavior affects handover timing.

Finally, co-manipulation tasks involve both agents manipulating the same object, usually over an extended period of time. Such tasks include manipulating both rigid and deformable objects. For example, Mortl et al. [48] explored role allocation in dyads moving a large table, while Koustoumpardis et al. [33] investigated a control system for collaborative fabric folding. These tasks typically involve force sensing, as people communicate through haptic channels during co-manipulation [54].

In addition to the type of task, the task space may range from discrete to continuous. The way in which a researcher chooses to view the task space will likely depend on the research question, and it will affect the analysis methods the researcher can use. For example, a researcher could view a shared workspace task as a series of discrete events. In this case, they may analyze the interaction based on the timing of human and robot actions or by counting the number of times a certain event occurs (for example, see [29]). Alternatively, they could consider it as a continuous interaction, in which case they might analyze the data by looking at, say, the trajectories of a robot arm and human arm, or by exploring causal relationships between their actions.

The research question will also influence the task context the researcher chooses. For instance, if the researcher wants to explore a technical research question about a new control architecture for handovers, the task context will naturally be a handover. However, other researcher questions may be answerable in different task contexts, and it may even be interesting to explore them in more than one task context. For example, an interaction-focused research question about proxemics could be explored in any of the three task contexts mentioned above, and it could be informative to see how the relative positions of the human and robot change in different task contexts. Also, the platforms available to a researcher may influence which task context they choose (discussed in Section 3.3), and these decisions may be made in tandem.

The task context can affect many choices downstream, including sensing requirements and analysis techniques. For instance, some implementations of shared workspace tasks or handovers require the robot to recognize the state of the task, or track precise movements of the person or robot in real time, which can require sophisticated sensor systems, such as motion capture. This may not be required in co-manipulation tasks, where the human and robot often have a static grip on the object. Additionally, in contrast to shared workspace tasks and handovers, several common fluency metrics, such as idle time and concurrent activity, are not applicable in co-manipulation tasks. Because the agents are both continuously active during the task, such measures are uninformative. However, because the agents are physically coupled for the entire task, other metrics, such as interaction forces, become available. Thus, the type of task chosen will impact the analysis techniques available to pxHRT researchers.

3.3 Platform

Researchers must make many decisions concerning the platform(s) to use in pxHRT experiments. A wide variety of robotic systems are available with varying capabilities, morphologies, and prices. For instance, researchers must choose whether they want to use a mobile robot, such as the PR2 or Pepper [21, 56, 70], or a stationary robot, such as a Sawyer arm or Jibo.

Different robots also have a wide variety of manipulation capabilities. Many social robots, like Kuri, have no manipulators, but may provide affordances for teaming which may not be as readily apparent in a Fetch or PR2. On the other hand, robots without manipulators can limit pxHRT research to either mobility or language-related questions.

In addition to considering the tasks a given number of manipulators can perform, pxHRT researchers should also consider the complexity of the system. The more manipulators a platform has, the harder it becomes to avoid collisions. Furthermore, as the number of degrees of freedom of the manipulator(s) increases, motion planning and control becomes more difficult.

Other capabilities of the robot must also be considered. The safety features, accuracy, computation speed, maximum velocity, and strength of the robot will affect the types of tasks the robot can do. For example, Baxter robots are considered relatively safe because their joints are somewhat compliant, but they have a positional accuracy of ± 5mm. In contrast, the Kuka KR 60-3 has a pose repeatability of ± 0.06mm, but is not compliant. Thus, a researcher interested in, say, the co-manipulation of objects in home settings could use a Baxter, but someone focused on collaborative precision manufacturing might prefer a Kuka.

The morphology of a robot can also affect people's interaction with it [9]. Thus, whether to use a humanoid robot like an iCub, a zoomorphic robot like an Aibo or Paro, or a mechanistic robot like a TurtleBot should be reflected on, especially in more social settings.

Due to the costs associated with robot hardware, researchers may be restricted to use whatever robots that they have at hand and/or are inexpensive. This can affect the tasks contexts they are able to explore.

3.4 Sensors

In a proximate HRT study, researchers potentially need to sense many things, including the human, robot, environment, objects, and task state. Sensors vary in their sensing modalities, accuracy, and reliability, particularly when it comes to sensing human and robot motion [34]. Like platforms, cost can affect which sensors a researcher uses; often they use what is already available. Thus, for resource-limited laboratories, it can be helpful to both consider data collection and analysis techniques which are flexible to work well with lower-cost sensing systems (e.g., Kinect vs. MoCap), against the precision required for the method.

There are many ways to sense people during interaction. Some sensors focus on specific parts of the body, such as arms or legs, while others sense the entire body. For example, an IMU sensor might just sense a particular limb, while a motion capture system could be used to sense the whole body. Therefore, a researcher exploring handover tasks might want to only monitor the person's arm, in which case an IMU sensor on the arm might be adequate and much cheaper than a motion capture system. Furthermore, some sensors are better suited for detecting gross motion, while others can detect fine motion [34]. Therefore, someone interested in co-manipulation of bulky objects might opt for sensors that are better for gross motion, whereas if their focus is on handovers, they might be better off with a sensor that more precisely identifies fine motion. Additionally, some sensors, like an accelerometer, can detect motion, while others, such as a heart rate monitor or galvanic skin response sensor, collect physiological signals.

Researchers also often need a way to sense the robot they use, both in terms of monitoring for errors and tracking the physical state of the robot. For instance, it will be necessary to monitor robot errors that could affect the experiment, such as a controller crashing or the robot performing an erroneous action. Software failures are often tracked in log files. Additionally, if the robot acts incorrectly, this can often be identified manually, either by watching during the experiment or looking at video recordings afterward. However, depending on the length and style of the experiment, this can be time-consuming.

It may also be helpful to know the joint positions and location of the robot. This can often be done easily with encoders the robot already has, making it a convenient sensing modality. In some studies, such as a shared workspace study with the person and robot working in close proximity, it may be necessary to track the robot's location in a room to determine its location relative to users. In some cases, this could be done by having the robot keep track of its position in a map, a capability many robots possess. However, while this is often relatively straightforward and inexpensive, it can sometimes result in unacceptably large dead reckoning errors. Because of this, it may be necessary in some instances to track the robot with a motion capture or other sensing system.

Sensing the environment is also important, both for the robot's function and for data analysis. The robot may need to avoid obstacles in its environment or determine the current state of the task. This can sometimes be done with sensors on the robot. For instance, the Toyota HSR has a LIDAR sensor that it can use to avoid obstacles. Researchers will also need to track the environment to determine the success of the interaction. For instance, in a co-manipulation task, they may want to determine task progress or how precisely a human and robot place an object in a given position. Researchers can use a variety of sensors to sense the environment, including LIDAR, cameras, microphones, and force sensors.

What a researcher chooses to sense and the types of sensors used will affect the degree of autonomy the robot can have during the study. For instance, if the robot can not sense its position in space or sense objects around it, it may not be able to be fully autonomous in a co-manipulation task. Similarly, if it cannot sense the position of a person's hand, it may be difficult for it to be autonomous in a handover task. Sensors will also affect the analysis methods available to a researcher. For example, to evaluate the proxemics of a human-robot team, the researcher must have some way to measure the position of both agents. To analyze the relative timing of events, the researcher needs a way to sense when events occur.

3.5 Autonomy

The autonomy of a system is an important factor in a pxHRT experimental design. A robot may range from being fully teleoperated to being fully autonomous. Some systems are partially teleoperated, with a person manually controlling one or more aspects of its behavior, such as speech generation, or the person may share control of the robot with an autonomous agent. Furthermore, some systems react dynamically to the people around them or the environment, while in others, the actions are predetermined and static. Autonomy is a complicated subject, and a full discussion is beyond the scope of this paper; for a more complete consideration of autonomy in HRI, see [5].

There are different reasons a researcher might choose to use a teleoperated system or an autonomous system. For instance, it may be easier to teleoperate the system. In other cases, the experimenter may initially teleoperate a system to collect data to train an autonomous system on. Conversely, for some research, a teleoperated system may not be an option. If the researcher is examining the efficacy of a controller they designed, they would not be able to use a fully teleoperated system. Additionally, there are some methodological and ethical concerns with the Wizard of Oz (WOZ) paradigm of teleoperation [61].

If the researcher chooses to use a semi- or fully-autonomous system, they will likely need a way to sense and model the components of their study. This may involve sensing or predicting the person's intention, trajectory, or preferences. They will need to either sense or know a priori what state the environment is in. For instance, if the human and robot are moving around the room carrying an object, the robot might need a map of the room the experiment takes place in. Furthermore, the researcher will need to have some way to model the task the human and robot perform, as well as a way to monitor the state the task is in. Finally, they will need to make decisions about how to model the robot itself. This could involve the low-level controller used to control the robot's joints or a high-level control architecture.

The level of autonomy will also affect the evaluation in a number of ways. For one, some evaluative paradigms may be excluded by the level of autonomy chosen. For instance, it may be difficult for a person to teleoperate a robot that is expected to consistently respond to multiple types of social signals from multiple people at once.

Additionally, the autonomy will influence the assumptions researchers make in their evaluative paradigm. These will vary from study to study, but could include assumptions about knowledge of the world state [53] or the observability of human actions or task state [52]. Researchers may also assume prior knowledge of the task domain [70] or exclude certain aspects of the task [41]. These assumptions, as well as the level of autonomy in and of itself, will affect the ecological validity of the study and should be stated when reporting the study.

3.6 Evaluation

Researchers will want to select an evaluative paradigm, including the specifics of the tasks the team will perform; the different conditions, if any, in the study; the number of humans and robots per trial; where the study will take place; and so on. The details of the study evaluation should be determined before collecting data for the study to reduce ethical concerns and increase the validity of the study [68].

In general, in study design, there are many decisions researchers must make. Researchers will determine the specifics of the task in the experiment, such as how many times tasks should be performed. They will also decide where the study will take place and how long it will last. Furthermore, they will set the conditions of the study, and participants should be randomly assigned to these conditions.

Researchers also need to make decisions about participants, such as whether to use a between-subjects or within-subjects design. Furthermore, researchers should consider the population participants are drawn from and how many participants to use. To determine the sample size, researchers may conduct a power analysis. This requires an estimation of the anticipated effect size in the study (which can sometimes be estimated from prior work in the area) and the statistical power [26]. Sometimes, researchers also determine the number of participants based on what is typical in the field, but power analyses are recommended.

Researchers will also need to determine how many people should participate in a given trial of the study. If they are interested in how groups of people interact with a robot, they will need at least two (and possibly more) participants in each trial [21, 29, 50]. On the other hand, if they are interested in dyadic interactions between a person and robot, they will only need one participant per trial [53, 59]. For example, if researchers want to explore how a robot can adapt to a person's level of fatigue in a dyadic interaction, they will need one participant per trial [59]. Other times, the appropriate number of human participants may be less obvious. For instance, if the research question concerns people's social interactions with a robot, the way they interact might change based on whether they are alone with the robot or also with other people. In this case, researchers

must decide which case they are most interested in studying (or include this as a variable in their study).

Regardless of the population or number of participants, researchers should also be aware of and plan for effects on participant behavior from the conduction of the study. For instance, people sometimes change their behavior when they believe they are being observed, which is known as the Hawthorne effect [6]. They may interact differently with the robot in the context of an experiment than they would outside that context. In-the-wild studies where people do not know they are being observed can minimize the Hawthorne effect, as well as deceiving people about the nature of the study (with proper IRB approval and informing them about the deception after the study) [6]. Aside from this, the experimenter conducting the study may also unintentionally give participants cues about what is expected from them, which could change the results [26]. These effects can be managed in a variety of ways, including keeping experimenters blind to the condition they administer, preventing the participant from seeing the researcher during the experiment, and telling participants that the experimenter does not have personal involvement with what is being investigated in the study [26].

Additionally, when including human subjects in an experiment, researchers will typically need approval from an ethics review committee, known as an institutional review board (IRB) in the United States. The purpose of IRBs is to ensure that participants are treated ethically. Most studies involving human subjects need approval from an IRB before running the study, though if the experiment involves minimal risk to subjects, the approval process may be expedited. When running a study, make sure to plan for the time it will take for the appropriate ethics committees to review it.

Finally, researchers must make decisions about the number of robots and the specifics of the robot(s) behavior in their study. Similarly to the human subjects, researchers will decide how many robots to use in each trial of the study. While at least one robot is often involved, in some cases, it may be informative to see how people interact without a robot present in some conditions (e.g., [35, 60, 73]). For example, researchers may be interested in comparing how two people collaborate with each other using only haptic communication, vs. a person and a robot [60]. Researchers will also need to decide on the specifics of the robot's behavior. What exactly will the robot be doing during the task, and how will it do this? This may be tied to the level of autonomy of the robot. For instance, if experimenters want to test an algorithm they developed, the robot will act based on that algorithm, at least in one condition of the study.

The evaluative paradigm chosen will affect which analysis methods are most appropriate. For example, techniques like cross correlation may require longer trial lengths than timing or accuracy metrics. The points in the experiment at which questionnaires are administered will affect the analyses that can be performed with those questionnaires. Also, as discussed above, decisions like whether the study is between- or within-subjects will affect which analysis techniques can be used.

3.7 Analysis Methods

Finally, researchers will want to select analysis methods that well-support their research question. This may involve evaluating the team, human, robot, or task. For instance, they might evaluate the degree to which the team had a shared model of the task [53] or the degree of comfort the human felt [73]. They could look at the amount of force on the robot's gripper or task completion time [25]. The components of their study that they want to analyze will affect the method they select.

Researchers can choose to analyze a wide variety of aspects of their study. They may want to analyze people's perceptions of and attitudes towards the robot. They could also investigate task performance, information exchange, agent relationships, and causal relationships.

For any given aspect of interest, there are often many different ways researchers can measure and analyze it. For instance, researchers could measure the fluency of an interaction based on the idle time during a task, by functional delay, or both [23]. Furthermore, most measurements and analyses (with some notable exceptions, such as validated questionnaires) can be used to determine different things. For example, researchers might analyze the distance between a robot and people while a robot navigates a hallway, which could indicate how safe the robot is. However, another study might use that same measure (e.g., distance) as an indication of how comfortable people feel around the robot (if they are more comfortable with the robot, they might be willing to get closer to it). In this subsection, we will discuss a variety of constructs researchers may want to investigate in studies, and metrics and analyses for these constructs. We attempt to include a broad array of methods, but this overview is not comprehensive: there are many constructs of potential interest to researchers, and many ways to analyze them.

Additionally, the type of data they collect can be quantitative or qualitative (or both). Qualitative data is often in the form of language [26]. It is often analyzed using techniques like thematic coding. Meanwhile, quantitative data can somehow be translated into a number and is often analyzed using null-hypothesis significance testing [6, 26]. However, there are concerns with this method, particularly because p-values are often very different between experiments and because the α values used to determine significance are proscribed somewhat arbitrarily. Furthermore, because researchers have many choices when analyzing data, they may unconsciously find a significant result even when no actual effect exists [19]. Additionally, reporting results as "significant" based on the p-value being below a threshold sometimes misleads people to believe the difference is meaningful, even though it may or may not be. Thus, some researchers suggest adopting other types of statistical tests (such as Bayesian statistics) or for reporting confidence intervals rather than p-values [6, 10, 15].

3.7.1 Perceptions and Attitudes. People's perceptions and attitudes towards a robot (or robots) are often of interest to pxHRT researchers. For example, researchers might want to know how comfortable people feel around the robot [73], how much they trust the robot [53], or how fluent they perceived the interaction to be [23]. One way to gauge people's perceptions is by asking them, such as through interviews or a questionnaire. Interviews can allow researchers to gain insight into people's thoughts. To analyze interviews, they may perform thematic analysis on the interviews, finding common themes across participants' responses. When conducting interviews, interviewers should be careful not to ask leading questions. Otherwise, the results may be influenced by the interviewer's underlying expectations or biases, rather than reflecting the participant's thoughts.

Questionnaires are a common and straightforward way to evaluate many aspects of teaming, and can be used to evaluate human perceptions of workload, fluency, comfort, trust, usability, preferences, and so on [23]. For instance, questionnaires indicated that participants were less comfortable with a robotic assistant when its lights were flashing [73]. Nikolaidis and Shah [53] used a questionnaire to find that participants trusted their system more compared to a baseline, and Cakmak et al. [11] used a forced choice question to determine which robot poses people interpreted as indicating a handover. There are many different types of questionnaires available to researchers, and sometimes researchers develop their own questionnaires when no existing ones explore the concepts they are interested in. In general, researchers should try to use a validated questionnaire when possible, and they should validate any questionnaire they develop [64]. Likert scales are one commonly used type of questionnaire, and there are many validated Likert scales. However, recent work suggests that Likert scales in particular are commonly misused in HRI research [64], so researchers should be mindful when using them to follow best practices. For instance, when

analyzing data from a scale, the entire scale should be analyzed, not just data from a single item in the scale.

Outside of directly asking people about their perceptions and attitudes, researchers can also attempt to infer them from people's behaviors. This can take many different forms, from measuring the distance between a person and robot [53], to measuring physiological signals like skin conductance [2]. Researchers should be careful to ensure that the behavior they measure is actually related to the perception or attitude they are interested in. Often, researchers use both self-reported measures and observational or sensor-measured data (e.g., both questionnaires or interviews and behavioral data) to assess people's perceptions and attitudes.

3.7.2 Task Performance. Furthermore, experimenters might want to examine task performance. There are a variety of ways in which they can do this. One common measure is how long it takes a team or individual to complete a task [16, 21, 25, 41, 48, 52, 60]. Accuracy is another common metric to measure task performance. The accuracy of a system can also indicate the performance of an algorithm or control paradigm [16, 46, 56]. Additionally, it can be used to evaluate how well a person and robot work together [16]. Generally, lower task completion times and higher accuracy are considered better, though this may be complicated by factors like human preference (e.g., [7]).

Task performance can also be described by counting the number of times an event occurs. For instance, Talamadupula et al. [70] looked at the number of resource conflicts in a simulation when a robot attempted to coordinate a plan with a human.

Various timing metrics can also be used to characterize task performance. For instance, to evaluate a system's efficacy in a group task a researcher might examine how appropriate the robot's timing is in a group context [29]. Another metric of potential interest for characterizing task performance is burstiness, which describes the distribution pattern of an activity and can range between periodic, random, and very bursty [84]. Hoffman [23] also proposed several task fluency metrics that rely on the timing of actions between robots and humans, specifically functional delay, human and robot idle time, and the amount of time spent acting concurrently.

Researchers may also want to compare individual task performance to team task performance, or human-human performance with human-robot performance. This will often involve similar measures to the ones to characterize task performance generally, but will compare differences between types of teams. For example, Lamb et al. [35] counted the number of times an avatar or human picked up an object and the percentage of passes from one to the other to evaluate the similarity between a human-human team and a human-avatar team. Reed et al. [60] looked at how long an individual took to complete a task, as compared to a dyad, to investigate differences between human-human and human-robot collaboration with only a haptic communication channel. Similarity measures can also help researchers determine how similar their robot's actions are to a human's, as well as the balance of the partnership between the robot and human. Iqbal et al. [29] used the group synchronization index to determine how closely a robot timed its moves with a human group.

3.7.3 Information Exchange. Researchers may be interested in information exchange between agents performing the task. They may use accuracy or the time it takes a person to react to a robot during a task to determine how well a person interprets a robot's intent [17]. Nikolaidis and Shah [53] investigated how well a team converged on a shared policy for a task by looking at the similarity of the action sequences the human and robot expected. Researchers can also examine the information exchanged haptically between agents [42, 60].

3.7.4 Agent Relationships. Another potential aspect of interest is the relationships between the agents. This could include the roles that agents take on during a task, as well as how well they

coordinate or work together. As discussed in Section 2, there are many possible ways to characterize and analyze a relationship, including stability, mode, and adaptation. For instance, Reed et al. [60] used the degree of specialization in a team to characterize different team types and roles. Morl et al. [48] examined role allocation and how such allocation affected people's perception of the experience.

Force measurements are one way to examine relationships between agents. These can be used to characterize interactions by looking at the disagreement between partners [48] or the contribution of participants [60]. Mortl et al. [48] used the magnitude of force that agents exerted against each other during a collaborative table moving task to evaluate the extent the agents disagreed with each other during the task. Reed et al. [60] looked at the force profiles from two agents in a collaborative crank turning task to determine how much each agent contributed.

3.7.5 Causal Relationships. Causal relationships are another way to characterize relationships, both between team members and between other variables in the study. Windowed cross-correlation and Granger Causality are two techniques that give insight into such relationships. For example, Xu et al. [84] provide an example of Granger Causality between various parent and infant actions while playing with toys (e.g., parent manual activity, infant gaze, etc.). Medina et al. [41] used Granger Causality to determine the relationship between variables in a human-human handovers. Papaioannou et al. [56] looked at the causality between the duration of a conversation between a human and robot and the tasks they completed.

Finally, researchers may gain insights through qualitative analysis of their data, such as through the use of visualization techniques. This approach is commonly used when analyzing time series interaction data [75, 84]. Patterns they visually notice can indicate which quantitative analyses might be most useful in future studies, and can be informative in and of themselves (e.g., [30]).

Data analysis can be an iterative process. The initial techniques used to analyze the data may not provide meaningful results, in which case, researchers might explore other techniques. Additionally, results from one analysis might point to other analyses that could be more informative. Researchers can then include these analyses in future studies. However, when reporting results from a study, it is important to explicitly note which analyses were planned and which are exploratory to avoid p-hacking [26, 68].

4 CONSTRAINTS

There are many constraints that must be considered when designing a study. Time and monetary constraints unavoidably must be dealt with. This includes the amount of time and money required to design and obtain the required components of the study, run the study, and analyze the resulting data. The complexity of the components involved will affect the time and cost. While this framework will not eliminate these constraints, we hope that using it will assist researchers in fully using their time and resources in a thoughtful, structured manner.

There are also many challenges that will likely arise during a study. These include hardware-, systems-, and human-related challenges, as well as data labelling and analysis challenges.

Hardware inevitably fails occasionally, and researchers must deal with the outcomes. For instance, Hoffman and Breazeal [25] experienced a hardware failure that made them unable to run five participants, as well as two minor failures from which they were able to recover. However, in spite of these setbacks, they were both able to produce meaningful results from the other participants in their study. Therefore, if time and resources allow, it is often better to schedule more participants than might be strictly necessary for the study in case of failures like these. When considering what sensors and analysis techniques to use, it is generally advisable to account for possible failures.

Systems-related challenges are also common. The dialogue system employed by Papaioannou et al. [56], for example, was not completely adjusted to their shopping mall setting, which sometimes resulted in inappropriate or unhelpful answers. Sometimes it is possible to mitigate such system limitations. For instance, many researchers limit users to a small subset of phrases when communicating with the robot, which increases the accuracy of speech recognition systems. Some people also choose to use a WOZ paradigm for aspects of the robot's behavior that are difficult to implement, especially if those components are not the main focus of the study.

Another challenge with many systems in HRT is that they have a high degree of complexity, and often have many components. For instance, Talamadupula et al. [70] attempted to integrate four subcomponents into their system, and were not able to fully include one of them. Additionally, some systems require large amounts of data to train on, which is not always readily available. Because of this, researchers sometimes use simulators to generate more data. For instance, Lombardi et al. [36] used simulated agents to model people to create a data set on which to train their system.

Including participants in a study also comes with its own innate challenges. Sometimes participants are unable to complete tasks or perform tasks incorrectly [48, 73]. Participants had so much difficulty completing the last segment of a table moving task that Mortl et al. [48] discarded that portion of the data. Fortunately, these problems can often be somewhat mitigated by looking at other segments of the task, as Mortl et al. [48] did, as well as by using many sensing modalities and employing a variety of analysis techniques.

Another common research challenge involves data labelling and analysis. Some types of data need to be coded by hand, which requires extra resources. For instance, Unhelkar et al. [73] had two raters independently code their data. If a researcher does not have the resources to code their data, they may prefer to use techniques that do not require hand-coding. Other times, researchers collect too much data to process in a reasonable amount of time. For example, Medina et al. [41] decided to simplify their grip force measurement by summing all of the force measurements from the hand. If more data is collected than can be analyzed, it is often possible to analyze a subset of the data, for instance, to analyze just the head movements from a motion capture system. In this case, the excess data can be analyzed in later studies. However, researchers need to be careful when doing this because choosing what data to analyze or report after running the study can result in higher rates of false positives [68]. Whenever possible, researchers should conduct and report all the analyses they planned in advance, and if they are not able to do so, they should clearly state this and explain which analyses they chose to conduct and why.

Researchers will not be able to foresee every snag that will occur during their study, but they can prepare in advance for some that are likely to occur. Additionally, some challenges, such as limited resources for data analysis or integrating pieces into a system, can be mitigated by careful planning. Researchers can decide what things they will need to do and estimate how much time each task will take them, reducing the chance that they will run out of time for any given piece of the study.

Researchers may face a variety of challenges when running a study. Of course, this framework will not make such problems disappear. However, by helping researchers thoroughly plan their study in advance and prepare for such potential difficulties, we hope that they can be mitigated and that dealing with them will be easier.

5 PILOT STUDY DEMONSTRATING AN APPLICATION OF THE FRAMEWORK

In our research program, we are exploring the relationship between workload and teaming dynamics in human-robot dyads, with a particular focus on mobile co-manipulation in safety critical domains. We now walk through a case study from our own work to show how we used our proposed framework to guide us. While we explain our study design linearly for clarity, most decisions were made in parallel to some extent. For instance, when deciding whether our task context should be

Fig. 3. Tasks from our experiment. Our tasks included co-manipulation with deformable and rigid objects and a shared workspace task, as well as a handover.

discrete or continuous, we also considered what kinds of analysis methods we might want to use, since those decisions affect each other.

We began by deciding on a research question. To determine what types of questions we might want to ask, we began by reviewing relevant literature in the field. To date, much work on pxHRT has focused on shared workspace tasks (e.g., [13, 21, 24, 53, 73]), as well as handovers [11, 69] . Even in these spaces, standard metrics were only recently proposed. While many studies employed measures such as the magnitude of concurrent human and robot activity, human and robot idle time, and functional delay, these were only evaluated in 2019 [23]. Additionally, they were evaluated in, and have primarily been used in, shared workspace tasks, though some studies have also used them to investigate handovers [11, 28].

Comparatively little has been done to evaluate people's perceptions of coordination during human-robot co-manipulation tasks. Faria et al. [17] investigated coordination between multiple people and a robot when a robot poured water into cups the people held. They analyzed the interaction using functional delay and reaction time, as well as self-reported measures. Additionally, Nikolaidis et al. [51] evaluated a table carrying task, but only used self-reported measures for coordination. The study was also online, which detracts from its ecological validity with respect to real-world, physical human-robot interactions. Therefore, we were curious about what kinds of coordination patterns people showed when performing co-manipulation tasks with a robot.

We also noticed that most studies focus on a single kind type of task, but in an environment like a home, a robot may be expected to perform many different types of tasks, and smoothly switch between them. Therefore we were interested in how people coordinate with robots while switching between several tasks in a continuous interaction.

Additionally, we were interested in how workload might affect people's coordination with and reliance on a robotic partner. Prior work indicates that people might rely more on automation in a high workload condition [81]. If people consistently perform differently under high workload compared to low workload, robots may be able to use this information to make better inferences about a person's mental state or what is happening in the environment.

Thus, we were interested in the movement dynamics that humans exhibit with robots in different task contexts under different workloads. One research question was whether mental workload changes a human-robot team's fluency while performing a variety of pxHRT tasks. We hypothesized that the fluency and people's perceptions of fluency would be different based on workload for all types of tasks, as people may rely on the robot differently between low- and high-workload conditions. Another question was how mental workload affects human-robot temporal coordination patterns, particularly during co-manipulation tasks. We expected that the person might lead the

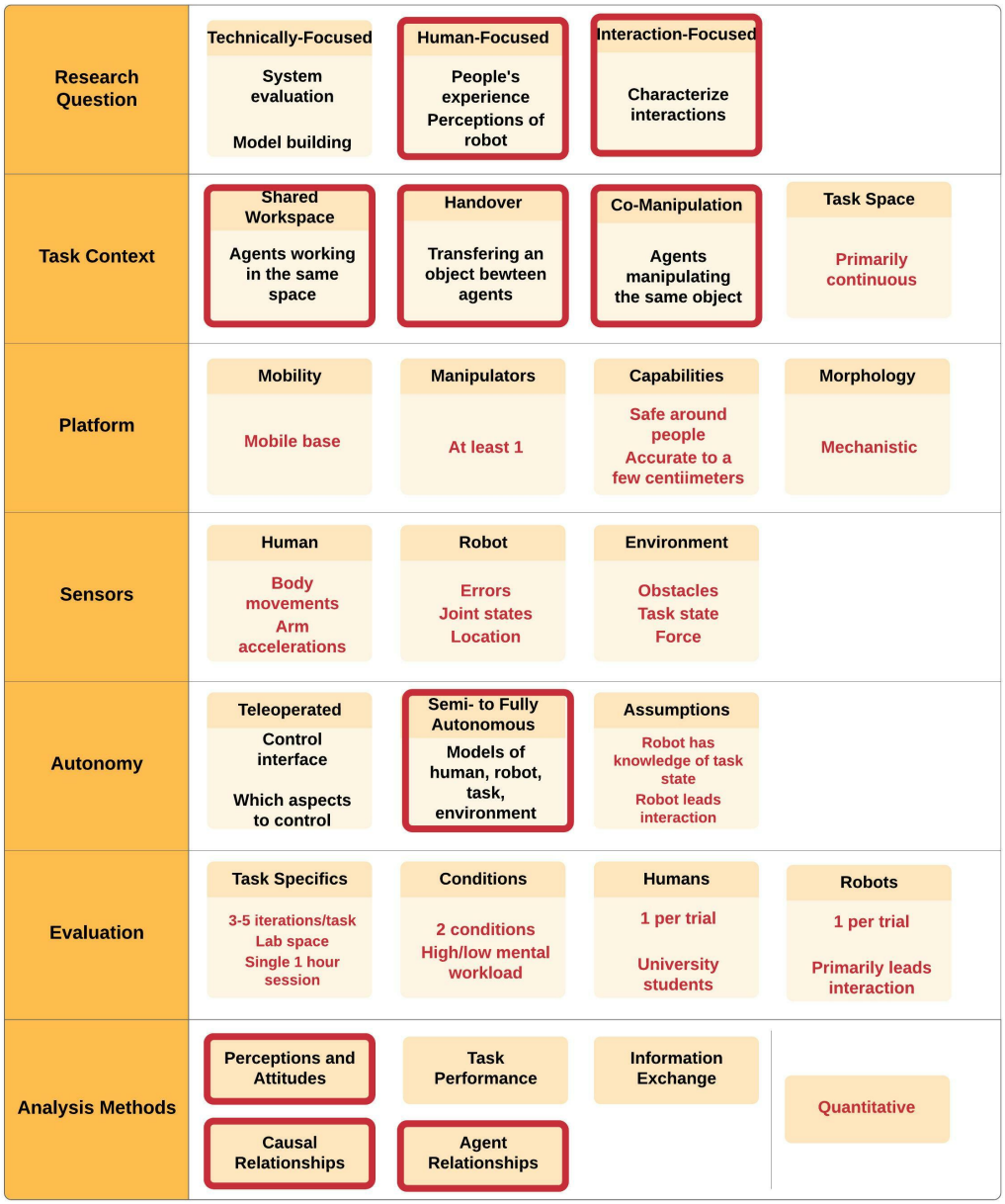

Fig. 4. The decisions we made for the pilot study based on our framework. Options we chose are in red.

interaction more in a low-workload condition and less in a high-workload condition. We chose interaction- and human-focused questions because we felt we needed to develop a method to characterize human-robot teaming in different task contexts before developing a control paradigm for a robot-focused study, and we felt interaction- and human-focused studies are more suitable for the development of such methods. In future work, we plan to incorporate the results of this

human-focused study into a control architecture for proximate teaming, and then test this control architecture in a technically-focused study.

We then needed to select a task context. Because we were interested in the dynamics between a person and robot performing different tasks during a continuous interaction, we needed to include different types of tasks. We decided to include tasks from the three different task types identified by Hoffman [23]: co-manipulation, handovers, and shared workspace tasks. When thinking about the potential tasks we could include, we also considered that co-manipulating rigid and deformable objects might be different, and thus included both in our study. Specifically, the person and robot co-manipulated sheets and boxes, first placing sheets over a couch, and then moving boxes around the room. Then, they alternated between handing over a paint roller and painting a stencil together. By including these different kinds of tasks, we could explore whether the dynamics expressed by the person and the fluency changed based on the task context.

Additionally, we had to decide whether we would view the task space as discrete or continuous. We chose to primarily view the space as continuous, since many coordination analyses use continuous data. However, we also decided to include some discrete data, such as when the robot started a task and when a handover was initiated and finished. This would allow us to use some of the fluency metrics from [23], such as functional delay.

We also chose a platform to use. In order to perform co-manipulation tasks and handovers, the robot needed at least one manipulator. We also wanted the robot to be mobile in order to perform a co-manipulation task that involved motion around the room. Furthermore, we thought about some of the capabilities the robot would need to complete the tasks in the study. It needed to be relatively safe around people, since people would closely interact with it during the handovers and shared workspace task. The robot needed to be somewhat precise to complete the tasks we had in mind, though most tasks could still be completed if the robot had inaccuracies of several centimeters. Additionally, we would like the robot to be able to move at a typical walking pace to facilitate moving around the room with a person while co-manipulating an object. We were not particularly concerned about the morphology of the robot, as how the morphology affected people's coordination strategies was not the focus of our study, though it could be an interesting investigation in the future. With these considerations in mind, we chose to use the Toyota HSR, as it was the only robot readily available to us that had both a mobile base and a manipulator. It also fulfilled most of our other requirements, as it was designed to be safe around people and would be precise enough to complete the study tasks. It was not quite as fast as we would have liked, but we felt this limitation was acceptable, since people often adapt their walking pace to others.

With this platform in mind, we needed to choose what sensors to use. Since our research question involved the person and robot's interactions, we needed to sense both of their actions. We were particularly interested in the person's motion patterns, so we needed ways to sense their movements. However, we did not have a motion capture system available to us, so instead we opted to use skeleton tracking with a Kinect, a much cheaper option. The Kinect video data also allowed us to keep track of the objects in the environment, as well as certain robot failures, such as when the robot attempted to pick up a box and missed. We used Myo armbands, which collect IMU and sEMG data, on both arms of participants. We also used the built in force sensor in the robot's gripper to measure the amount of force the person applied on the robot during some of the tasks.

Because our research questions involved human-robot interactions, we needed to sense the robot in addition to the human. To sense the robot, we primarily employed sensors the robot was already equipped with, such as encoders in the joints. We also made a map of the lab space and used the robot's mapping capabilities to track its location in the room. These techniques to sense the robot were easy to implement and simple to collect data from, factors important to us due to time constraints and the number of people on the project.

During the experiment, the robot also needed information about its environment if we wanted it to run autonomously. We used the force sensor in the gripper to provide information about the person and certain events in the environment, and we used the Lidar to help localize the robot. For instance, during the shared workspace task, the robot used its force sensor to determine when it had come into contact with the paper it was supposed to be painting.

We also needed to decide on the level of autonomy of the robot. We chose to have the robot be fully autonomous, as it would be difficult for a person to consistently and repeatably teleoperate the robot for multiple 15 minute trials. Additionally, having an autonomous system increased the ecological validity of the study, as all of the robots actions were within its actual capabilities without human assistance. We decided to have the robot lead the interaction, as we could still address our research questions this way, and it did not require an in depth model of the person. However, during co-manipulation tasks, the robot assumed that when there was an upward or downward force on its gripper (minus an approximation of the force of gravity), it meant the person wanted to move the box up or down. We thought this would provide an easy way for people to influence the interaction. We also provided the robot with full knowledge of the task and task state, as we assumed that in most cases where it led the interaction, it would have access to this information.

Additionally, we decided on the specific evaluative framework to employ. Due to the specific networking requirements of the robot we used, we set this study up in a lab space. This decreased the ecological validity of the study, and to partially make up for this, we wanted a coherent narrative to tie the tasks together. Therefore, we framed all of the tasks as pieces of a collaborative painting scenario. In this framing, the person and robot first performed a co-manipulation task with sheets (a non-rigid object) to cover a couch, as if to protect it during painting. Then, they moved several boxes away from the painting area, and moved some boxes with supplies to the painting area (a co-manipulation task with a rigid object). Finally, the person put paint on a roller, handed the roller to the robot, held a stencil while the robot painted it, and took the roller back from the robot (alternating between handover and shared workspace tasks).

Because many coordination analyses perform better with multiple tasks, we decided to have the person and robot repeat each task multiple times in each trial. Because the robot was autonomous, we were able to ensure that it did each task repeatably (outside of failure cases). For example, the person and robot carried five sheets together, one after another. We also randomized the order of the workload conditions, since we were manipulating that variable.

To answer our research question about the effects of workload on fluency, we needed to manipulate participants' workload. We chose to have two conditions: a high mental workload and a low mental workload condition. In the high mental workload condition, participants needed to count the number of times they heard a certain tone amid other irrelevant tones, similarly to [8]. Aside from this, the high mental workload condition was the same as the low mental workload condition.

We were interested in dyadic interactions, and so chose to have one human participate with one robot in each trial. This was a pilot study, so we used a convenience sample. We recruited university students via word of mouth, and scheduled anyone willing to participate, resulting in 14 participants. This is not a particularly deliberate way to determine the sample size, but it is a simple way to recruit participants for a pilot study. We also decided to do a within-subjects study design, as nothing in our design prevented this, and it provides greater statistical power [26].

We also needed to determine the specific robot behaviors during the task. We decided to have the robot lead most of the interaction, as this made programming the robot a bit easier, but would still allow us to address our research questions. We also included some pieces that the person could lead, which we thought might make the task feel more collaborative. Specifically, the person could decide how high the box or sheet was during the co-manipulation task. If they raised or lowered

the object, the robot would sense the change in force in its gripper and would respond accordingly. The person also could adjust the position of the robot before placing the sheet on the couch.

Finally, we chose the analysis techniques to use. We needed to assess people's perceptions of workload and fluency in order to address our research questions. We decided to use validated survey measures to evaluate the workload [22] and fluency [23] experienced by the human, as these were easy to administer and provided quantitative data relevant to our questions. We also decided to use functional delay as an objective measure of fluency, as it is fairly simple to implement and has been shown to coincide with people's perceptions of fluency [23]. However, we only implemented this for the handover task, as it did not make sense for the co-manipulation task. During the shared workspace task, it was not clear at what point the person began experiencing delay, as they often adjusted the stencil throughout the task. With further study, using these self-reported survey measures along with more sensor-measured data could allow us to examine the connections between these different types of metrics.

For our second research question, we also needed to analyze the person's and robot's movement patterns during the co-manipulation task. We wanted to examine causal and agent relationships, as these can be informative in regards to how agents coordinate with each other. Therefore, we decided to analyze the co-manipulation tasks using Granger Causality to see which aspects of the interaction could be used to predict other components. We chose this method because it indicates if one time series is dependent on another time series, which can be used to infer agent relationships.

We also conducted cross-correlation and lag analysis on these tasks to see which agent led the other during the interaction. This indicates if one time series follows another, and by how much it lags behind the other. Granger Causality and cross-correlation are both common techniques for analyzing coordination between agents.

While we are also interested in using other complex systems analyses, such as cross-recurrence quantification analysis, these are more difficult to implement. Therefore, due to time and resource constraints, we chose to save those for future work. We had also planned to use the Kinect data to analyze the proxemics displayed by the person during the shared workspace task. However, again due to resource constraints, we leave this to future work.

5.1 Procedure

Our study was approved by our University's Institutional Review Board, under protocol number 161808XX.

Upon arriving at the lab, participants were asked to sign consent forms and read instructions for the study. After this, they put one Myo armband on each arm and confirmed their understanding of the task. They were able to ask one of the researchers questions at any time during the experiment.

The experiment consisted of two phases. In one phase, the person only did the tasks with the robot. In the other, the person performed the tasks with the robot while doing a cognitive loading task. Specifically, they had to count the number of times they heard a certain auditory stimulus among other irrelevant noises, as in [8]. The conditions were counterbalanced. After each phase, the participant completed the NASA Workload TLX scale [22] and the seven sub-scales for assessing fluency proposed by Hoffman [23].

Within each phase, the participants first moved five sheets with the robot to cover a couch. This provided a co-manipulation task with a deformable object. After moving all of the sheets, the person and robot moved boxes together. This task enabled us to examine a co-manipulation task with a rigid object. During these tasks, if the person tried to move the sheet or box up or down, the robot would reactively move its arm up or down in response to the change in force. After completing the co-manipulation tasks, the person and robot painted together in a task that combined a shared workspace task and a handover. The person held a series of three stencils on a piece of paper, and

the robot painted each stencil in with a paint roller (see Fig. 3). In between stencils and after the last one, the robot handed the person the paint roller to apply more paint or to return the paint roller to the tray. We analyzed the co-manipulation tasks with Granger Causality and cross-correlation, and the handover task with functional delay.

Due to time constraints we leave analysis of the shared workspace task for future work.

5.2 Results

In this section, we will discuss results from running the pilot study. While our sample size of 14 participants would typically be too small to show statistically significant results (unless there is a large effect size) or to generalize to the rest of the population, we include such results here for demonstration. We follow the templates from [26] and [18] for reporting statistical tests in Section 5.2.3.

5.2.1 Participants. A total of 14 people were opportunistically recruited by word-of-mouth to participate in our experiment. 6 were women and 8 were men. 12 were right-handed, 1 was left handed, and 1 was ambidextrous. Their ages ranged from 21 to 30 years old, with a mean of 23.4 years. They self-reported an average mid-level experience with robots, and the majority were graduate students. They were randomly assigned to do the high cognitive workload task in the first phase or second phase of the experiment. Participants were compensated with $10 gift cards. Data from three participants during the box carrying task and one participant during the painting task was discarded due to missing data points.

5.2.2 Manipulation Check. To make sure our manipulation had the desired effect, we performed a manipulation check using people's responses to the NASA Workload TLX scale. To check our manipulation, we ran a Wilcoxon signed-rank test. We found that participants reported a higher mental demand in the high workload condition ($M = 3.2$, $SD = 0.9$) than in the low workload condition ($M = 1.6$, $SD = 1.0$), $W = 99.5$, $p = 0.001$. This indicates that our cognitive task was effective at increasing people's experience of mental demand. No other measures from the Workload TLX were significantly different ($p \geq 0.45$), suggesting that the cognitive task was specific to mental load and that the human-robot collaboration was not perceived as more physically or temporally demanding while performing a cognitive task.

5.2.3 Analysis. Overall, our quantitative analyses were not as informative for these tasks as we expected. For instance, to check if there was a difference in functional delay during handovers under different workload conditions, we checked the data distribution for uniformity and determined we needed a non-parametric test. Therefore, we ran a Wilcoxon signed-rank test. We found that there was no significant difference in functional delay between the high mental workload ($M = 5.2$, $Std = 4.7$) and low mental workload ($M = 4.5$, $Std = 2.3$), $W = 51$, $p = 0.74$. This suggests that the cognitive task did not interfere with the handover. It is possible that the handover employed in our study was simple enough that cognitive load did not affect it.

Additionally, we ran a Granger causality analysis between accelerations of the human's arm and robot's base and arm, using the Multivariate Granger Causality Toolbox in Matlab [4]. This test also showed no statistically significant causal relationships for either covering the couch with sheets ($p > 0.07$) or carrying boxes ($p > 0.90$). We had expected that robot acceleration would be Granger Caused by the human arm acceleration in the box carrying task because the robot was programmed to move its arm up or down if it felt a change in the force on its gripper through the box. However, during the experiment, relatively few participants took advantage of being able to move the box up and down, with several of our participants running the box into a table rather

than lifting it over the table. This could be due to the participants forgetting that they were able to move the box vertically, as discussed in Section 5.3.

We also found the data for the maximum lag between conditions for the sheet and box tasks was non-parametric, so we ran a Wilcoxon signed-rank test. For the sheet task, we found no significant difference between the high workload ($M = 9.3$, $Std = 17.8$) and low workload ($M = -3.8$, $Std = 11.2$) conditions, $W = 53$, $p = 0.08$. Similarly, for the box task, there was no significant difference between the high workload ($M = 4.0$, $Std = 16.7$) and low workload ($M = -3.1$, $Std = 9.4$) conditions, $W = 51$, $p = 0.12$.

We ran a Wilcoxon signed-rank test to check if any of the self-reported measures of fluency in Hoffman's scale [23] were significantly different between conditions. We found that scores on subscale 5, which indicates team improvement, were significantly lower in the high mental workload condition ($M = 4.2$, $Std = 1.1$) than in the low mental workload condition ($M = 4.6$, $Std = 1.1$), $W = 1.5$, $p = 0.02$. No other subscales showed significant differences. This is interesting given that the robot was not learning and did not change its behavior during the study. This could reflect that the person was able to concentrate more on collaborating with the robot in the low workload condition. It is possible that this allowed the person to improve at the task, thus improving team performance.

5.2.4 Exploratory Analyses. To look at trends that might inform future study iterations, we performed some exploratory, qualitative analyses. We initially plotted our raw data, but we realized the data from our sensors was a bit noisy, making it hard to visually observe patterns. Therefore, we filtered the human right arm acceleration and robot arm acceleration with a fourth-order, low-pass Butterworth filter. Examining these graphs suggested patterns in our data that our quantitative analyses could not capture. For example, during the sheet carrying task, several participants had a similar pattern in which the changes in acceleration of their arm appeared lower as the robot turned than during the rest of the task (see Fig. 5). Reviewing the videos of these participants, they appeared to stand still while the robot rotated around them.

Additionally, for most of the task, particularly before lifting the sheet over the couch, the robot's arm did not frequently accelerate, and the acceleration changes were small, despite large changes in the human's arm acceleration. In contrast, during the box carrying task, the robot's arm frequently changed in acceleration, and the acceleration changes appear larger in magnitude than during the sheet carrying task, even though changes in the human's arm acceleration were often smaller in magnitude than during the sheet carrying task (see Fig. 5). This suggests that the frequency and magnitude of acceleration changes could inform a robot about the rigidity of an object it co-manipulates with a person.

Though we have not yet had the opportunity to investigate the relative position and motion of the person's body and robot's base, we plan to analyze this data in the future, as this could be more informative than just looking at their arms.

5.3 Example Pilot Study Discussion

While our pilot study involved too few participants to yield substantial results, it revealed some trends that warrant further study. For instance, the person's arm tended to have smaller accelerations when the robot turned during the sheet task. In this interaction, the robot acted as a leader, but if this trend is also shown to occur when the person leads, it could be used as a signal to help the robot determine when the person wants to initiate a turn.

Additionally, the robot's arm tended to have larger accelerations and accelerate more frequently when co-manipulating a box with a person as opposed to a sheet. This was surprising given that most people had larger right arm accelerations for the sheet than the box, and the robot used the

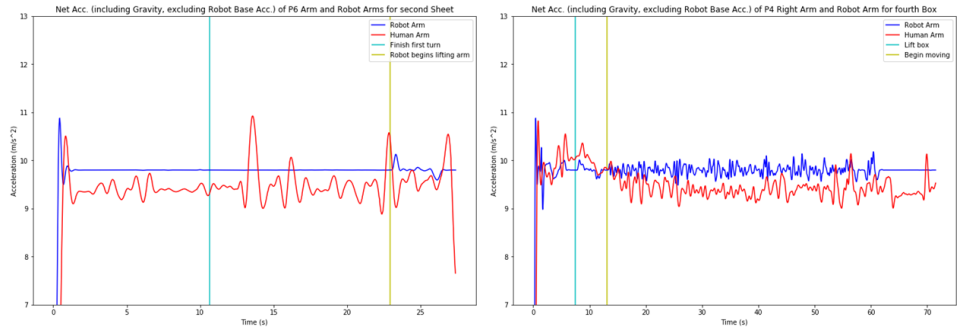

Fig. 5. Plots of the participant arm acceleration and robot arm acceleration for one round of the sheet carrying task (left) and one round of the box carrying task (right). Acceleration data was filtered with a fourth-order, low-pass Butterworth filter. The robot arm accelerates much less frequently and with a smaller magnitude during the sheet carrying task than during the box carrying task.

same control algorithm for both tasks. If this trend is confirmed in larger studies, it could be a useful way for robots to distinguish between rigid and non-rigid objects during co-manipulation tasks. This could be useful in enabling the robot to identify the object or determine the person's goal.

During the study, we also noted that our long-duration study design meant participants sometimes forgot which tasks they were supposed to perform, or what their role was. For example, many participants were unsure whether they were supposed to hand the paint roller or the stencil to the robot in the last task. This confusion likely impacted our results, as it resulted in participants interrupting their collaboration with the robot to ask the experimenter questions.

Our case study is intended to serve as an exemplar for our framework, and well-illustrates the complexity of conducting pxHRT studies.

6 CASE STUDY TO EVALUATE THE FRAMEWORK'S UTILITY

To determine the utility of our proposed framework, we also conducted a case study with researchers new to the field of pxHRT. Observing how researchers without familiarity with the field used the framework provided us with insight into how others could best use it. Additionally, the participants gave us feedback which could be used in future pxHRT frameworks.

We recruited two groups of students who had not previously designed or run pxHRT studies and had them design a pxHRT study using the framework. Both groups were in the process of planning pxHRT related studies when we recruited them. Group 1 consisted of three students who had recently completed their undergraduate degrees in scientific fields (P1, P2, and P3). They were planning a study to investigate how a child could remotely teleoperate a mobile manipulator to interact with other children. These students had a general idea of the questions they wanted to ask, the task context for the experiment, and the robot they planned to use, but they had not yet considered many details of the study.

Group 2 consisted of two engineering Master's students (P4 and P5). They were planning a study to look at how a robot running a navigation algorithm from our lab [72] could be situated in an emergency department, possibly to deliver supplies to healthcare workers. Their study design was a bit further along than Group 1, as they had discussed some ideas with a Ph.D. student who was leading the project.

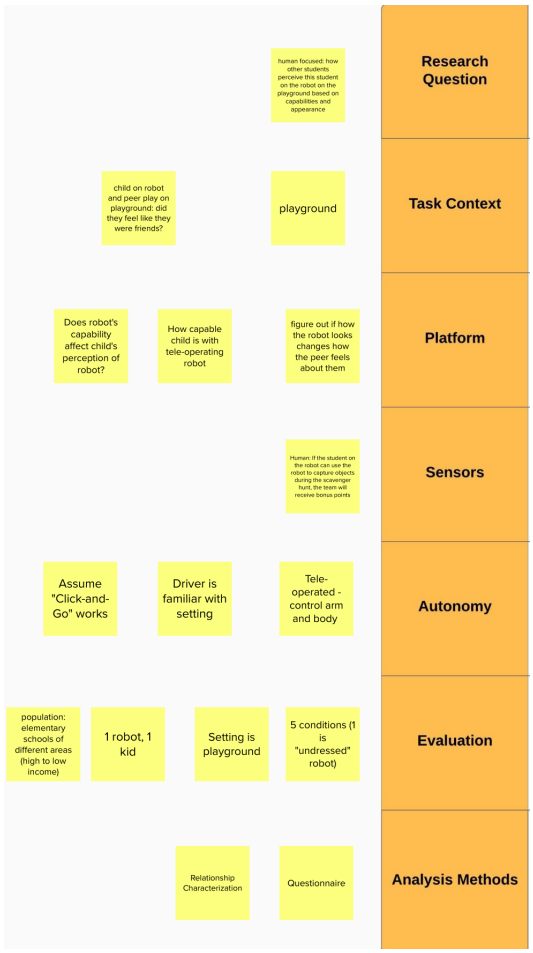

Fig. 6. The Mural created by Group 1 while designing their study.

Both groups of students participated in a one hour study session over Zoom. During the sessions, we first provided a brief overview of the framework and its components, as well as a walk through of the example pilot study to demonstrate how the framework could be used. The students then worked together to brainstorm a study design using the framework in Mural [1]. The Mural included images of Fig. 1, Fig. 2, and bullet points of the choices from the pilot study for the students to reference. They were asked to try to describe what they were thinking aloud as they worked.

Throughout the process, they were allowed to ask the experimenter any questions. After finishing their study design, they briefly explained their study design to the experimenter, and then participated in a semi-structured interview to understand participants' thoughts about the framework and give them an opportunity to provide feedback.

We recorded the sessions and reviewed the footage to see the participants' design process. We also conducted thematic analysis on the interviews and found common themes in participants' responses, which we report below.

While designing their study, Group 1 primarily used Fig. 2 to facilitate their design. They mostly discussed one component at a time, but sometimes jumped between different components. For

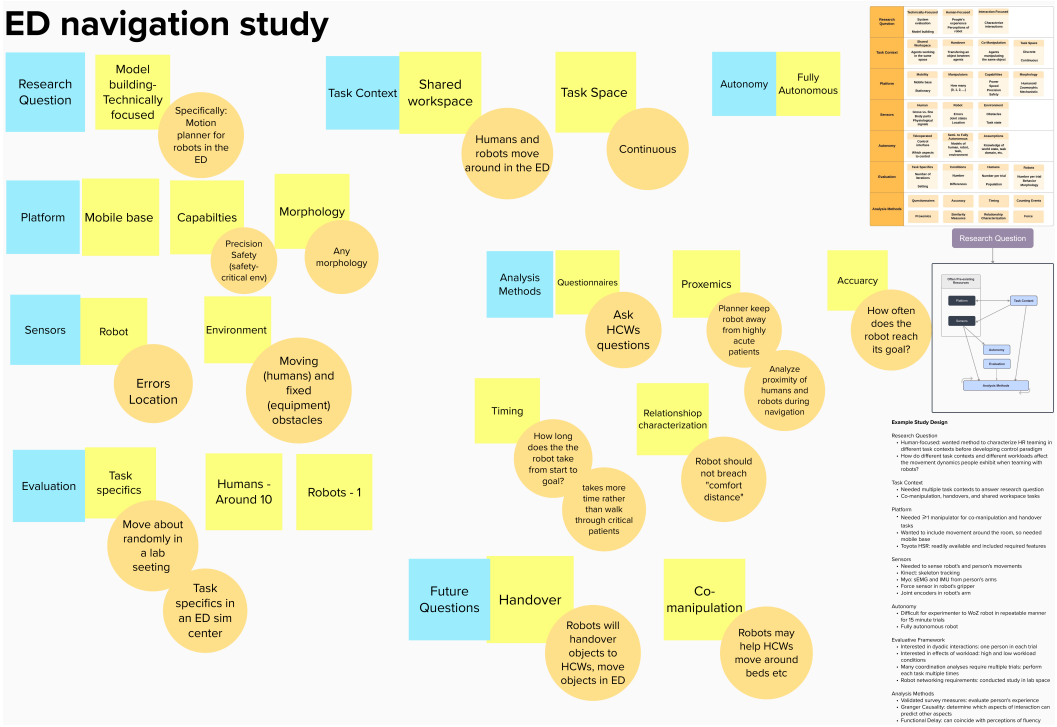

Fig. 7. The Mural created by Group 2 while designing their study.

instance, they started by specifying the type of research question they wanted to ask and what in particular they wanted to investigate, and then moved onto the task context. However, while discussing the task context, they also started talking about pieces of the evaluation, including the setting, conditions, and participants, as well as the robot morphology. Additionally, while discussing the sensors, they also considered how the abilities of a person operating the robot might change based on the sensors, and how this might affect the interactant's perception of the operator and robot, which was relevant to their research question.

As mentioned above, the students in Group 2 were further along in planning their study than the students in Group 1. They went through the framework more linearly than Group 1. While outlining their study, they realized that there were some questions they were interested in, but could not address in their current study. Therefore, in addition to the components of the framework, they included a section for future questions, where they listed ideas they would like to explore in the future.

From the semi-structured interviews, we learned that many of the students felt that the *framework provided structure during the study design process.* For instance, P4 noted that "it made it really easy to plug in the different things that we really wanted under different categories." P1 said that they were "walking through the different boxes in the framework" and were "trying to check those boxes off." P5 thought that "it's good because it kind of helps to narrow down things and gives us targets."

Additionally, several of the students felt that it helped them *focus on parts of the study they had not previously considered* and consider the details of the study design in more depth. P2 observed that "we've never talked about like how many kids are going to be in the group or like how we're going

to evaluate it [...] I think like reading through the framework like really helped us like think about that and [...] we kind of just like never considered that before." Similarly, P1 said, "We would've probably just come up with more surface level idea instead of making it more flushed out."

Interestingly, some students also thought that the framework could help them in ways we had not considered. P4 and P5 thought it might have assisted in their literature review process by enabling them to better target their review. They posited that they might have been able to look for specific literature in each component of the framework after outlining their study, thus narrowing down their search. Additionally, P1 and P3 thought the framework could help them explore a greater variety of possibilities. P1 felt it "helps you [...] think of things that you wouldn't have maybe explored before." P3 said that it "helped me generate new ideas and questions" and "think more broadly."

Some students also expressed some concerns about the framework. P2 reflected that she "didn't really think beyond [the categories presented in the framework]" and wondered if she could be "missing something." P2 and P3 also felt that the version of Fig. 1 we provided them with was a bit confusing, and P3 was not sure where to start with it. To alleviate this, we added a more thorough description of our framework and the connections between components at the beginning of Sec. 3.

Furthermore, participants provided suggestions to potentially improve the framework. P4 felt that safety was a potentially important topic that was not covered, and thought analyses for safety should potentially be included in the analysis methods. P4 and P5 also recommended adding a place for future research questions in the framework, which could be useful in projects with multiple stages or to break down a research question into smaller pieces. Additionally, P1 suggested adding "guiding questions" that could help researchers use the framework.

From our case study, we observed that researchers new to the field of pxHRT were able to use our framework to design a study. Even participants who had not read the paper and received only a brief introduction to the framework seemed to be able to utilize it well. Furthermore, the participants felt that the framework was useful, providing structure during their design process and enabling them to consider more details in their design. This indicates the potential utility of our framework.

7 DISCUSSION

pxHRT is an important environment to consider within HRI, as many key applications require robots to act contingently, fluently, and safely around people. Therefore, it is imperative that we have well-designed, well-tested systems, and strategies for human-robot teaming.

Furthermore, many of the spaces robots are expected to enter, and are currently being deployed in, are safety critical, such as hospitals or disaster zones. In these spaces, it is especially important to minimize risk and promote good teaming, as the consequences of mistakes can include injury or death. Thus, we need well-adapted teaming strategies for these spaces, and we need to develop them quickly, as robots are already moving into such environments.

However, pxHRT research is complex. In any given experiment, there are likely to be several sensors, one or more robots, and one or more humans. These systems are all often individually complex, and the interactions between them only add to the complexity.

As participants in our case study indicated, our proposed framework provides pxHRT researchers with a structured method to consider tradeoffs associated with experiments in proximate teaming, and could help researchers new to the field plan studies in more detail or consider new aspects of study design. By providing researchers with a structure to carefully plan studies, we hope to improve the quality of study design, which may produce results with more impactful implications. This in turn will allow for quicker advancements in pxHRT research and better-designed teaming systems.

Aside from the practical considerations of the field, from a more theoretical perspective, pxHRT is a cognitively complicated space because it combines concepts from a wide range of fields, from control theory to human factors to haptics. It is not possible for any one person to have a well-rounded background in all of these fields. However, pxHRT researchers need to be aware of techniques from these different spaces in order to explore new applications for existing methods, as well as to design new ones. Our proposed framework provides researchers with a structure to better understand their work within the context of pxHRT, as well as possible directions they can explore.

With this framework, we hope to enable pxHRT researchers to more thoroughly consider the components of their studies, more easily design experiments, and more fully explore pxHRT.

8 ACKNOWLEDGMENTS

We thank the Toyota Research Institute and Air Force Office of Scientific Research (AFOSR) for supporting this work under Grant No. FA9550-18-1-0125.

REFERENCES

[1] [n.d.]. Mural. https://www.mural.co/

[2] Kumar Akash, Wan-Lin Hu, Neera Jain, and Tahira Reid. 2018. A Classification Model for Sensing Human Trust in Machines Using EEG and GSR. ACM Transactions on Interactive Intelligent Systems 8, 4 (nov 2018), 27:1–27:20. https://doi.org/10.1145/3132743

[3] Kim Baraka, Patrícia Alves-Oliveira, and Tiago Ribeiro. 2020. An extended framework for characterizing social robots. In Human-Robot Interaction. Springer, 21–64.

[4] Lionel Barnett and Anil K. Seth. 2014. The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference. Journal of Neuroscience Methods 223 (2014), 50–68. https://doi.org/10.1016/j.jneumeth.2013.10.018

[5] Jenay M. Beer, Arthur D. Fisk, and Wendy A. Rogers. 2014. Toward a framework for levels of robot autonomy in human-robot interaction. Journal of human-robot interaction 3, 2 (2014), 74–99.

[6] Tony Belpaeme. 2020. Advice to New Human-Robot Interaction Researchers. In Human-Robot Interaction: Evaluation Methods and Their Standardization. Springer, Cham, 355–369. https://doi.org/10.1007/978-3-030-42307-0_14

[7] Tapomayukh Bhattacharjee, Ethan K. Gordon, Rosario Scalise, Maria E. Cabrera, Anat Caspi, Maya Cakmak, and Siddhartha S. Srinivasa. 2020. Is More Autonomy Always Better? Exploring Preferences of Users with Mobility Impairments in Robot-assisted Feeding. In Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. ACM, New York, NY, USA, 181–190. https://doi.org/10.1145/3319502.3374818

[8] Megan J. Blakely, Simon Kemp, and William S. Helton. 2016. Volitional Running and Tone Counting: The Impact of Cognitive Load on Running Over Natural Terrain. IIE Transactions on Occupational Ergonomics and Human Factors 4, 2-3 (2016), 104–114. https://doi.org/10.1080/21577323.2015.1055864 arXiv:https://doi.org/10.1080/21577323.2015.1055864

[9] Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. 2016. Social robotics. In Springer handbook of robotics. Springer, 1935–1972.

[10] Frank Broz, Chris S. Crawford, Hatice Gunes, Astrid Rosenthal-von der Putten, Laurel Riek, and Megan Strait. [n.d.]. Reproducibility in Human-Robot Interaction: Furthering the Science of HRI. In review ([n. d.]).

[11] Maya Cakmak, Siddhartha S. Srinivasa, Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2011. Using Spatial and Temporal Contrast for Fluent Robot-Human Hand-overs. In Proceedings of the 6th international conference on Human-robot interaction - HRI '11. ACM Press, Lausanne, Switzerland, 489–496. https://doi.org/10.1145/1957656.1957823

[12] Andrew Chang, Steven R. Livingstone, Dan J. Bosnyak, and Laurel J. Trainor. 2017. Body sway reflects leadership in joint music performance. Proceedings of the National Academy of Sciences 114, 21 (may 2017), E4134–E4141. https://doi.org/10.1073/PNAS.1617657114

[13] Crystal Chao and Andrea Lockerd Thomaz. 2016. Timed Petri nets for fluent turn-taking over multimodal interaction resources in human-robot collaboration. The International Journal of Robotics Research 35, 11 (sep 2016), 1330–1353. https://doi.org/10.1177/0278364915627291

[14] Tiffany L Chen, Tapomayukh Bhattacharjee, J Lucas McKay, Jacquelyn E Borinski, Madeleine E Hackney, Lena H Ting, and Charles C Kemp. 2015. Evaluation by expert dancers of a robot that performs partnered stepping via haptic interaction. PloS one 10, 5 (2015), e0125179.

[15] Geoff Cumming. 2014. The New Statistics: Why and How. Psychological Science 25, 1 (nov 2014), 7–29. https://doi.org/10.1177/0956797613504966

[16] Joseph DelPreto and Daniela Rus. 2019. Sharing the Load: Human-Robot Team Lifting Using Muscle Activity. (2019). http://people.csail.mit.edu/delpreto/icra2019

[17] Miguel Faria, Rui Silva, Patricia Alves-Oliveira, Francisco S. Melo, and Ana Paiva. 2017. "Me and You Together" Movement Impact in Multi-user Collaboration Tasks. In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, Vancouver, BC, Canada, 2793–2798. https://doi.org/10.1109/IROS.2017.8206109

[18] Andy P. Field. 2003. How to design and report experiments. Sage publications Ltd., London.

[19] Andrew Gelman and Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Department of Statistics, Columbia University 348 (2013).

[20] Ali Ghadirzadeh, Judith Butepage, Atsuto Maki, Danica Kragic, and Marten Bjorkman. 2016. A Sensorimotor Reinforcement Learning Framework for Physical Human-Robot Interaction. In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, Daejeon, South Korea, 2682–2688. https://doi.org/10.1109/IROS.2016.7759417

[21] Matthew C. Gombolay, Reymundo A. Gutierrez, Shanelle G. Clarke, Giancarlo F. Sturla, and Julie A. Shah. 2015. Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. Autonomous Robots 39, 3 (oct 2015), 293–312. https://doi.org/10.1007/s10514-015-9457-9

[22] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Advances in psychology. Vol. 52. Elsevier, 139–183.

[23] Guy Hoffman. 2019. Evaluating Fluency in Human–Robot Collaboration. IEEE Transactions on Human-Machine Systems 49, 3 (jun 2019), 209–218. https://doi.org/10.1109/THMS.2019.2904558

[24] Guy Hoffman and Cynthia Breazeal. 2007. Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team. In Proceeding of the ACM/IEEE international conference on Human-robot interaction - HRI '07. ACM Press, Arlington, Virginia, USA, 1–8. https://doi.org/10.1145/1228716.1228718

[25] Guy Hoffman and Cynthia Breazeal. 2010. Effects of anticipatory perceptual simulation on practiced human-robot tasks. Autonomous Robots 28, 4 (may 2010), 403–423. https://doi.org/10.1007/s10514-009-9166-3

[26] Guy Hoffman and Xuan Zhao. 2020. A Primer for Conducting Experiments in Human-Robot Interaction. ACM Transactions on Human-Robot Interaction 10, 1 (oct 2020). https://doi.org/10.1145/3412374

[27] Nadine Homburg. 2018. How to Include Humanoid Robots into Experimental Research: A Multi-Step Approach. In Proceedings of the 51st Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences, 4423–4432. https://doi.org/10.24251/hicss.2018.559

[28] Chien-Ming Huang, Maya Cakmak, and Bilge Mutlu. 2015. Adaptive Coordination Strategies for Human-Robot Handovers.. In Robotics: science and systems, Vol. 11. Rome, Italy.

[29] Tariq Iqbal, Samantha Rack, and Laurel D. Riek. 2016. Movement Coordination in Human–Robot Teams: A Dynamical Systems Approach. IEEE Transactions on Robotics 32, 4 (aug 2016), 909–919. https://doi.org/10.1109/TRO.2016.2570240

[30] Tariq Iqbal and Laurel D. Riek. 2015. A Method for Automatic Detection of Psychomotor Entrainment. IEEE Transactions on Affective Computing 7, 1 (jun 2015), 3–16. https://doi.org/10.1109/TAFFC.2015.2445335

[31] Megan A Jacobs and Amanda L Graham. 2016. Iterative development and evaluation methods of mHealth behavior change interventions. Current Opinion in Psychology 9 (2016), 33–37.

[32] M. Kim, K. Oh, J. Choi, J. Jung, and Y. Kim. 2011. User-Centered HRI: HRI Research Methodology for Designers. Springer Netherlands, Dordrecht, 13–33. https://doi.org/10.1007/978-94-007-0582-1_2

[33] Panagiotis N. Koustoumpardis, Konstantinos I. Chatzilygeroudis, Aris I. Synodinos, and Nikos A. Aspragathos. 2016. Human Robot Collaboration for Folding Fabrics Based on Force/RGB-D Feedback. In Advances in Robot Design and Intelligent Control. Springer, Cham, 235–243. https://doi.org/10.1007/978-3-319-21290-6_24

[34] Alyssa Kubota, Tariq Iqbal, Julie A Shah, and Laurel D Riek. 2019. Activity recognition in manufacturing: The roles of motion capture and sEMG+ inertial wearables in detecting fine vs. gross motion. IEEE International Conference on Robotics and Automation (ICRA) (2019).

[35] Maurice Lamb, Patrick Nalepka, Rachel W. Kallen, Tamara Lorenz, Steven J. Harrison, Ali A. Minai, and Michael J. Richardson. 2019. A Hierarchical Behavioral Dynamic Approach for Naturally Adaptive Human-Agent Pick-and-Place Interactions. Complexity 2019 (jun 2019), 1–16. https://doi.org/10.1155/2019/5964632

[36] Maria Lombardi, Davide Liuzza, and Mario Di Bernardo. 2019. Deep learning control of artificial avatars in group coordination tasks. arXiv preprint arXiv:1906.04656 (2019). arXiv:1906.04656v1 https://arxiv.org/pdf/1906.04656.pdf

[37] Tamara Lorenz, Alexander Mörtl, and Sandra Hirche. 2013. Movement synchronization fails during non-adaptive human-robot interaction. In Proceedings of the 2013 ACM/IEEE International Conference on Human-Robot Interaction - HRI '13). IEEE, 189–190.

[38] Tamara Lorenz, Astrid Weiss, and Sandra Hirche. 2016. Synchrony and reciprocity: Key mechanisms for social companion robots in therapy and care. International Journal of Social Robotics 8, 1 (2016), 125–143.

[39] Ludovic Marin, Johann Issartel, and Thierry Chaminade. 2009. Interpersonal motor coordination: From human–human to human–robot interactions. Interaction Studies 10, 3 (jan 2009), 479–504. https://doi.org/10.1075/is.10.3.09mar

[40] Pedro Marques-Quinteiro, André Mata, Cláudia Simão, Rui Gaspar, and Ana Rita Farias. 2019. Observing Synchrony in Dyads Effects on Observers' Expectations and Intentions. Social Psychology 50, 3 (may 2019), 174–184. https://doi.org/10.1027/1864-9335/a000377

[41] Jose R. Medina, Felix Duvallet, Murali Karnam, and Aude Billard. 2016. A Human-Inspired Controller for Fluid Human-Robot Handovers. In 2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids). IEEE, Cancun, Mexico, 324–331. https://doi.org/10.1109/HUMANOIDS.2016.7803296

[42] Erich A. Mielke, Eric C. Townsend, and Marc D. Killpack. 2017. Analysis of Rigid Extended Object Co-Manipulation by Human Dyads: Lateral Movement Characterization. arXiv preprint arXiv:1702.00733 (2017).

[43] Lynden K. Miles, Jordan L. Griffiths, Michael J. Richardson, and C. Neil Macrae. 2010. Too late to coordinate: Contextual influences on behavioral synchrony. European Journal of Social Psychology 40, 1 (2010), 52–60. https://doi.org/10.1002/ejsp.721 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.721

[44] Lynden K Miles, Joanne Lumsden, Natasha Flannigan, Jamie S Allsop, and Dannette Marie. 2017. Coordination matters: interpersonal synchrony influences collaborative problem-solving. Psychology (2017).

[45] Lynden K. Miles, Joanne Lumsden, Michael J. Richardson, and C. Neil Macrae. 2011. Do birds of a feather move together? Group membership and behavioral synchrony. Experimental Brain Research 211, 3-4 (jun 2011), 495–503. https://doi.org/10.1007/s00221-011-2641-z

[46] Yasser Mohammad and Toyoaki Nishida. 2015. Learning interaction protocols by mimicking understanding and reproducing human interactive behavior. Pattern Recognition Letters 66 (nov 2015), 62–70. https://doi.org/10.1016/J.PATREC.2014.11.010

[47] AJung Moon, Daniel M. Troniak, Brian Gleeson, Matthew K.X.J. Pan, Minhua Zeng, Benjamin A. Blumer, Karon MacLean, and Elizabeth A. Croft. 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 - HRI '14. ACM Press, Bielefeld, Germany, 334–341. https://doi.org/10.1145/2559636.2559656

[48] Alexander Mörtl, Martin Lawitzky, Ayse Kucukyilmaz, Metin Sezgin, Cagatay Basdogan, and Sandra Hirche. 2012. The role of roles: Physical cooperation between humans and robots. The International Journal of Robotics Research 31, 13 (aug 2012), 1656–1674. https://doi.org/10.1177/0278364912455366

[49] Alexander Mörtl, Tamara Lorenz, and Sandra Hirche. 2014. Rhythm patterns interaction-synchronization behavior for human-robot joint action. PloS one 9, 4 (2014), e95195.

[50] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Footing In Human-Robot Conversations: How Robots Might Shape Participant Roles Using Gaze Cues. In Proceedings of the 4th ACM/IEEE international conference on Human robot interaction - HRI '09. ACM Press, La Jolla, California, USA, 61–68. https://doi.org/10.1145/1514095.1514109

[51] Stefanos Nikolaidis, David Hsu, and Siddhartha S. Srinivasa. 2017. Human-robot mutual adaptation in collaborative tasks: Models and experiments. International Journal of Robotics Research 36, 5-7 (2017), 618–634. https://doi.org/10.1177/0278364917690593

[52] Stefanos Nikolaidis, Ramya Ramakrishnan, Keren Gu, and Julie Shah. 2015. Efficient Model Learning from Joint-Action Demonstrations for Human-Robot Collaborative Tasks. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15. ACM Press, Portland, Oregon, USA, 189–196. https://doi.org/10.1145/2696454.2696455

[53] Stefanos Nikolaidis and Julie A. Shah. 2013. Human-robot cross-training: computational formulation, modeling and evaluation of a human team training strategy. In Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction - HRI '13. IEEE Press, 33–40. https://dl.acm.org/citation.cfm?id=2447563

[54] E. Noohi, M. Žefran, and J. L. Patton. 2016. A Model for Human–Human Collaborative Object Manipulation and Its Application to Human–Robot Interaction. IEEE Transactions on Robotics 32, 4 (Aug 2016), 880–896. https://doi.org/10.1109/TRO.2016.2572698

[55] Linda Onnasch and Eileen Roesler. 2020. A Taxonomy to Structure and Analyze Human–Robot Interaction. International Journal of Social Robotics 3 (jun 2020), 1–17. https://doi.org/10.1007/s12369-020-00666-5

[56] Ioannis Papaioannou, Christian Dondrup, Jekaterina Novikova, and Oliver Lemon. 2017. Hybrid Chat and Task Dialogue for More Engaging HRI Using Reinforcement Learning. In 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE, Lisbon, Portugal, 593–598. https://doi.org/10.1109/ROMAN.2017.8172363

[57] Priyam Parashar, Lindsay M. Sanneman, Julie A. Shah, and Henrik I. Christensen. 2019. A Taxonomy for Characterizing Modes of Interactions in Goal-driven, Human-robot Teams. In IEEE International Conference on Intelligent Robots and Systems. Institute of Electrical and Electronics Engineers Inc., Macau, China, 2213–2220. https://doi.org/10.1109/IROS40897.2019.8967974

[58] Luka Peternel, Nikos Tsagarakis, and Arash Ajoudani. 2017. A Human–Robot Co-Manipulation Approach Based on Human Sensorimotor Information. IEEE Transactions on Neural Systems and Rehabilitation Engineering 25, 7 (jul 2017), 811–822. https://doi.org/10.1109/TNSRE.2017.2694553

[59] Luka Peternel, Nikos Tsagarakis, Darwin G. Caldwell, and Arash Ajoudani. 2018. Robot adaptation to human physical fatigue in human-robot co-manipulation. Autonomous Robots 42 (2018), 1011–1021. https://doi.org/10.1007/s10514-017-9678-1

[60] Kyle B. Reed and Michael A. Peshkin. 2008. Physical Collaboration of Human-Human and Human-Robot Teams. IEEE Transactions on Haptics 1, 2 (jul 2008), 108–120. https://doi.org/10.1109/TOH.2008.13

[61] Laurel D Riek. 2012. Wizard of oz studies in hri: a systematic review and new reporting guidelines. Journal of Human-Robot Interaction 1, 1 (2012), 119–136.

[62] Leonel Rozo, Sylvain Calinon, Darwin G. Caldwell, Pablo Jimenez, and Carme Torras. 2016. Learning Physical Collaborative Robot Behaviors From Human Demonstrations. IEEE Transactions on Robotics 32, 3 (jun 2016), 513–527. https://doi.org/10.1109/TRO.2016.2540623

[63] Richard C Schmidt, Claudia Carello, and Michael T Turvey. 1990. Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. Journal of experimental psychology: human perception and performance 16, 2 (1990), 227.

[64] Mariah L. Schrum, Michael Johnson, Muyleng Ghuy, and Matthew C. Gombolay. 2020. Four years in review: Statistical practices of likert scales in human-robot interaction studies. In Proceedings of the ACM/IEEE International Conference

on Human-Robot Interaction - HRI '20. IEEE Computer Society, 43–52. https://doi.org/10.1145/3371382.3380739 arXiv:2001.03231

[65] Lee Sechrest and Aurelio José Figueredo. 1993. Program Evaluation. Annual Review of Psychology 44, 1 (1993), 645–674. https://doi.org/10.1146/annurev.ps.44.020193.003241 arXiv:https://doi.org/10.1146/annurev.ps.44.020193.003241 PMID: 19845457.

[66] Elisa S Shernoff, Ane M Maríñez-Lora, Stacy L Frazier, Lara J Jakobsons, Marc S Atkins, and Deborah Bonner. 2011. Teachers supporting teachers in urban schools: What iterative research designs can teach us. School psychology review 40, 4 (2011), 465–485.

[67] Kevin Shockley, Matthew Butwill, Joseph P. Zbilut, and Charles L. Webber. 2002. Cross recurrence quantification of coupled oscillators. Physics Letters A 305, 1-2 (nov 2002), 59–69. https://doi.org/10.1016/S0375-9601(02)01411-1

[68] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science 22, 11 (oct 2011), 1359–1366. https://doi.org/10.1177/0956797611417632

[69] Kyle Strabala, Min Kyung Lee, Anca D. Dragan, Jodi Forlizzi, Siddhartha S. Srinivasa, Maya Cakmak, and Vincenzo Micelli. 2013. Towards Seamless Human-Robot Handovers. Journal of Human-Robot Interaction 2, 1 (feb 2013), 112–132. https://doi.org/10.5898/JHRI.2.1.Strabala

[70] Kartik Talamadupula, Gordon Briggs, Tathagata Chakraborti, Matthias Scheutz, and Subbarao Kambhampati. 2014. Coordination in Human-Robot Teams Using Mental Modeling and Plan Recognition. In 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, Chicago, IL, USA, 2957–2962. https://doi.org/10.1109/IROS.2014.6942970

[71] Angelique Taylor, Darren M Chan, and Laurel D Riek. 2020. Robot-centric perception of human groups. ACM Transactions on Human-Robot Interaction (THRI) 9, 3 (2020), 1–21.

[72] Angelique M. Taylor, Sachiko Matsumoto, Wesley Xiao, and Laurel D. Riek. 2021. Social Navigation for Mobile Robots in the Emergency Department. IEEE International Conference on Robotics and Automation (ICRA) (2021).

[73] Vaibhav V. Unhelkar, Ho Chit Siu, and Julie A. Shah. 2014. Comparative Performance of Human and Mobile Robotic Assistants in Collaborative Fetch-and-Deliver Tasks. In 2014 9th ACM/IEEE International Conference on Human-Robot Interaction - HRI '14. IEEE, Bielefeld, Germany, 82–89.

[74] Marynel Vázquez, Aaron Steinfeld, and Scott E Hudson. 2015. Parallel detection of conversational groups of free-standing people and tracking of their lower-body orientation. In 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 3010–3017.

[75] James Walker, Rita Borgo, and Mark W Jones. 2015. TimeNotes: a study on effective chart visualization and interaction techniques for time-series data. IEEE transactions on visualization and computer graphics 22, 1 (2015), 549–558.

[76] Auriel Washburn, Akanimoh Adeleye, Thomas An, and Laurel D. Riek. 2020. Robot Errors in Proximate HRI: How Functionality Framing Affects Perceived Reliability and Trust. ACM Transactions on Human-Robot Interaction (THRI) 9, 3, Article 19 (May 2020), 21 pages. https://doi.org/10.1145/3380783

[77] Auriel Washburn, Mariana DeMarco, Simon de Vries, Kris Ariyabuddhiphongs, R. C. Schmidt, Michael J. Richardson, and Michael A. Riley. 2014. Dancers entrain more effectively than non-dancers to another actor's movements. Frontiers in Human Neuroscience 8 (oct 2014), 800. https://doi.org/10.3389/fnhum.2014.00800

[78] Charles Webber and Joseph Zbilut. 2005. Recurrence quantification analysis of nonlinear dynamical systems. Tutorials in Contemporary Nonlinear Methods for the Behavioral Sciences (01 2005).

[79] Charles L Webber Jr and Joseph P Zbilut. 2007. Recurrence quantifications: feature extractions from recurrence plots. International Journal of Bifurcation and Chaos 17, 10 (2007), 3467–3475.

[80] A. Weiss, D. Wurhofer, M. Lankes, and M. Tscheligi. 2009. Autonomous vs. tele-operated: How people perceive human-robot collaboration with HRP-2. In Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction - HRI '09. 257–258. https://doi.org/10.1145/1514095.1514164

[81] Christopher D. Wickens and Stephen R. Dixon. 2007. The benefits of imperfect diagnostic automation: a synthesis of the literature. Theoretical Issues in Ergonomics Science 8, 3 (2007), 201–212. https://doi.org/10.1080/14639220500370105

[82] Scott S. Wiltermuth and Chip Heath. 2009. Synchrony and Cooperation. Psychological Science 20, 1 (2009), 1–5.

[83] Alan M. Wing, Satoshi Endo, Adrian Bradbury, and Dirk Vorberg. 2014. Optimal feedback correction in string quartet synchronization. Journal of The Royal Society Interface 11, 93 (apr 2014), 20131125. https://doi.org/10.1098/rsif.2013.1125

[84] Tian Linger Xu, Kaya de Barbaro, Drew H Abney, and Ralf FA Cox. 2020. Finding structure in time: visualizing and analyzing behavioral time series. Frontiers in Psychology 11 (2020).

[85] Holly A. Yanco and Jill L. Drury. 2002. A Taxonomy for Human-Robot Interaction. In Proceedings of the AAAI Fall Symposium on Human-Robot Interaction. www.aaai.org

[86] Holly A. Yanco and Jill L. Drury. 2004. Classifying human-robot interaction: An updated taxonomy. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Vol. 3. The Hague, Netherlands, 2841–2846.