Excuse Me, Could You Please Assemble These Blocks For Me?

Andrew Silva[*]
andrew.silva@gatech.edu

Siddhartha Banerjee[*]
siddhartha.banerjee@gatech.edu

Sonia Chernova
chernova@cc.gatech.edu

ABSTRACT

We designed and conducted a user study in which we had to collect data, train a model, and analyze the effects of the model on specific performance metrics, all while obscuring the true nature of our study. We examine the importance of (1) model evaluation and selection, (2) proper participant motivation and instruction, and (3) active control of confounding factors. In this paper, we present an account of our experiences, some of our ad hoc solutions, and the lessons that we think are valuable for the HRI community.

1 INTRODUCTION

In this work, we examine the steps we took to design a user study for evaluating the effects of interruptibility-aware behavior in robots, and the setbacks that we encountered along the way. We begin by introducing the research questions that the study sought to answer. We then detail the final study procedure in order to provide a more specific context for our setbacks and lessons learned. Finally, we explore three key issues we encountered, along with insights for identifying and resolving those issues more generally.

Our research seeks to develop interruptibility-awareness in robots, and to evaluate the effects of this capability on human task performance, robot task performance, and on the human's interpretation of the robot's social aptitude. Specifically, we focus on the following research questions:

RQ1 Can an integrated system be developed to accurately estimate human interruptibility online on a robot platform?

RQ2 How does interruptibility-aware robot behavior affect human task performance when a robot regularly needs assistance?

RQ3 How does interruptibility-aware robot behavior affect robot task performance when relying on humans for assistance?

RQ4 Does a robot appear more socially adept if it interrupts humans at appropriate moments?

In order to evaluate these questions, we conducted a user study in which human participants took part in a mock manufacturing assembly activity. Participants were given their own construction tasks while a robot with tasks of its own would occasionally interrupt them to request assistance. The study was conducted between-subjects and had three conditions in which we varied the mechanism used by the robot to decide an appropriate moment to interrupt the participant.

*Random interruptions (**RND**).* the robot interrupted participants after it waited for a random amount of time, reflecting the current behavior of interruptibility unaware robots.

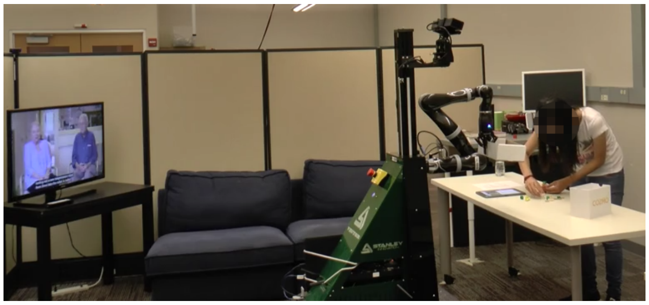

Figure 1: The robot interrupts a participant in a building task.

*Wizard-of-oz interruptions (**WOZ**).* the robot interrupted participants when a human (wizard) signaled it was an appropriate time. The wizard used the video stream from the robot's camera and was instructed to make moment-by-moment decisions on whether to interrupt.

*Intelligent interruptions (**INT**).* the robot interrupted participants based on output from an interruptibility classifier using a Latent-Dynamic Conditional Random Field (LDCRF) that we developed in prior work [1].

Our prior work had shown that the classification accuracy of the LDCRF was superior to other models in classifying interruptibility on a static dataset collected from our robot; we wanted to contextualize that success with an example of the model running online in a user study. To accomplish this goal, we used data from pilot studies and the RND condition to collect a dataset to train and test the model. Our self-contained and online interruptibility classification pipeline involved several state-of-the-art computer vision detectors, and is visualized in Fig 2.

2 FINAL STUDY DESIGN

The final study involved 48 participants recruited via email with 6 additional participants who took part in pilot trials used to tune build complexity, robot behavior, gather training data, and to familiarize the wizards with their interface. A previous iteration of the study involved 41 participants, though we had to abandon that design because it often encouraged undesirable participant behavior and yielded ineffectual data. Six of the final 48 trials were excluded from the study analysis: two due to hardware malfunction, and four

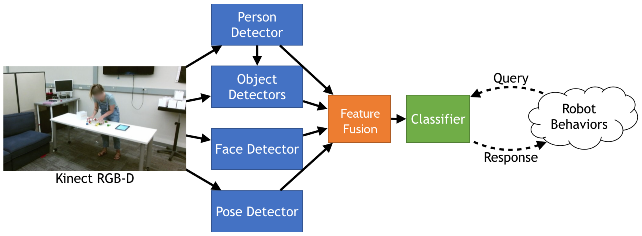

Figure 2: Perception and classification pipeline for interruptibility

[*] Authors contributed equally to this work

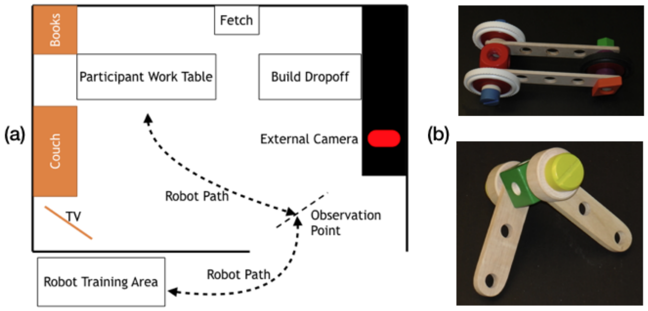

Figure 3: (a) Map of study area, (b), Sample builds for participants: (Top) Main Build, (Bottom) Interruption Build

due to participants deviating from the study protocol (Sec. 4). The resulting 42 participants (20 women, 22 men) were aged between 21 and 29 (*Mdn* = 24). Almost all participants were computer science students at Georgia Tech, with varying levels of robotics experience. The study took approximately 50 min, and all participants were paid $10 USD.

2.1 Study Procedure

In the experimental task, participants took part in a mock manufacturing assembly activity. Participants were instructed to construct structures (*builds*) out of wooden pieces (Fig. 3b), and told that their build process would be video recorded to be used later as training data for the robot. Additionally, participants were told that the robot was performing and studying its own builds, and that it would occasionally enter the space to request assistance.

Pre-Study. Upon arrival, participants were briefed on the study, completed consent forms, and filled in a pre-study questionnaire. Nearby, to support the narrative of the robot learning to construct builds, an experimenter could be seen "training" the robot by responding to the robot's questions (e.g., "Is this a correct build?").

Study Space. After the study briefing, participants entered the building area (Fig. 3a), consisting of an enclosed space with fetch area for retrieving build components, a work area for construction, and a dropoff area for completed builds. A key element of the study design is that the study schedule was split into periods of work and leisure to ensure that participants had periods of low and high interruptibility. To induce participants to showcase a diverse range of natural leisure behaviours (to fully evaluate the performance of the classifier and generalizability of our system), the room included a TV playing muted videos, a stack of books, and a couch. Participants were also allowed to keep their cell phones. Overall, during breaks 64% sat on the couch, 50% used their cell phones, 40% drank a refreshment, and 14% read a book.

For the remainder of the study period, participants alternated between constructing builds (*build*) and break times (*idle*), while being occasionally interrupted by the robot. Fig. 4 presents an example timeline.

Builds. Each participant trial consisted of 3 build sessions. The first build session was a training session during which participants were allowed to ask questions and acclimate themselves to the task and the robot. We do not report data from this session. Sessions 2 and 3 each consisted of two builds, with a short break in between. Instructions for each build were provided on a tablet located on the work table; the tablet remained blank until the designated build time, and presented a NASA-TLX workload questionnaire each time the participant selected that they had completed a build. The build sessions were either 15 min or 9 min in length, and were presented to all participants in a counterbalanced manner. The different length build sessions were configured to provide differing degrees of time pressure on the participant. In addition, pilot studies sometimes indicated a significant improvement in performance due to learning; the counterbalanced sessions were used to amortize any effects through the difference of learning during high time pressure and low time pressure sessions. All builds in a build session had a time limit, and participants were shown a countdown timer 30 sec before the end of this time limit; participants were not allowed to work past the end of the time limit. The tablet also presented the TLX questionnaire if participants ran out of time.

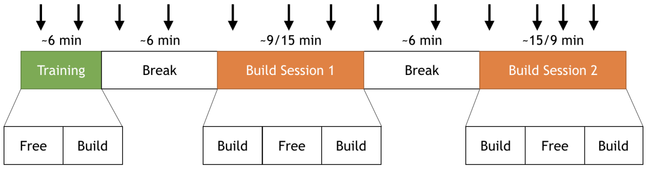

Figure 4: Sample timeline of a trial, arrows indicate interruptions

Breaks. Each trial included two break times approximately 6 min in length (differences in duration occurring due to robot interruptions), during which the tablet was taken away and the participants were invited to rest on the couch. The purpose of the break was to expose the robot to interruptible human behavior. In both cases, the experimenters presented fictitious excuses to the participant for pausing the study, in one case claiming a non-existent tracking device required adjustment, and in the other case simulating a tablet malfunction. For both breaks, experimenters explained the pause in the experiment, invited participants to wait on the couch, and then returned at the end of the break to "continue" the study. Participants were told that the robot interruptions would continue since the robot remained unaffected by the glitch.

Robot Interruptions. The robot continually entered the building area looking for assistance from the start to the end of a trial. The schedule of these entrances was not predefined and the robot was sent back in as soon as it returned from an interruption. The first three robot entrances coincided with the training build session and part of the first break; we allowed participants to ask questions during these interruptions and do not report data from them. The robot was equipped with a small box containing the blocks for its builds and a tablet, which provided instructions to the robot builds and presented a TLX questionnaire when done.

During an entrance, the robot followed the path shown in Fig. 3a. It waited at the observation point upon entering and after waiting—a random duration in RND, until an empirically chosen 2.5 sec of consecutive interruptible classifications in INT, or until the wizard sent an interruptible signal in WOZ—chose to move toward the participant. Upon arrival, the robot verbally requested assistance and waited for 2 min. Participants were aware of the wait duration and could accept the interruption within the time limit by grabbing

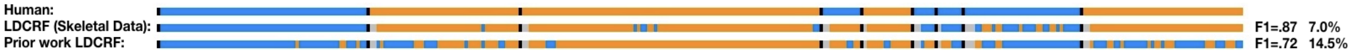

Figure 5: Classification timelines. We used the timelines to visualize both F1 score and a *fluctuation metric* (in percentage on the right)

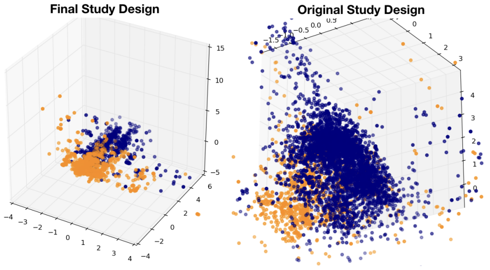

Figure 6: PCA of data from our final and original study design.

the tablet, at which point the robot waited indefinitely until the build was completed. If the participants did not respond in 2 min, the robot left the participant build area. Upon returning to the training area, the robot audibly requested verification of the build (e.g., "Is this a correct build?") from an experimenter. The experimenter provided a Yes/No response based on whether the interruption was built[1], prepared the next robot build, and sent the robot back in.

Post-Study. After the last build session, participants were asked to complete a post-study questionnaire, were debriefed on the purpose of the study and the deceptions that we used.

2.2 Metrics

In addition to logging data from the robot's cameras in the RND condition, we obtained quantitative measures of human and robot performance, and 5-point Likert scale responses to questions of participant opinions and participant background. The quantitative measures of human and robot task performance included metrics like: builds completed, total time building or idle, number of interruptions encountered and ignored, time taken to respond to the robot, time taken to complete robot tasks, and percent of appropriately- or inappropriately-timed interruptions.

Most quantitative measures were automatically logged from timestamps on the tablet and the robot, but some discrepancies caused by unexpected participant behavior[2] were corrected using video from the external camera. In addition to the above metrics, we also asked participants to verbally elaborate on their choices and reasoning during post-study debriefing.

3 LESSON 1: MODEL EVALUATION

Our work aimed to develop a fully autonomous and integrated interruptibility classification system. Despite our prior work's [1] success in effective interruptibility classification using accuracy scores, finding an appropriate model for use on the robot proved to be more complicated than simply comparing F1 scores. In this section, we highlight the issues we faced and our lessons from them.

3.1 Problems

Our training data for the model consisted of data from the robot while observing participants in the RND condition of both iterations of our study. We annotated each participant's moment-by-moment interruptibility and, following our established methods, trained LDCRF classifiers. However, despite seemingly high F1 scores, the model was not accurate when deployed in pilots, frequently oscillating between interruptibility classes for a seemingly static scene.

3.2 Solutions

Our primary solution was extensive visualization of the data and of our model predictions. As shown in Figs. 5 and 6, we made visualizers to (1) evaluate the features provided to the model with methods such as Principal Component Analysis (PCA) and (2) evaluate the consistency of model predictions with timelines.

Visualizers of the features in the data from the first iteration of our study showed us that our training data lacked diversity: uninterruptible participants exhibited a greater diversity in feature values post-PCA than interruptible participants. We were forced to draw the conclusion that our initial study design did not encourage participants to exhibit diverse leisure behaviours. As a result, we updated the study schedule to include forced breaks and extended periods of free time. We also made more distractor objects, such as the TV and the couch, available in the participants' workspace.

For evaluating and improving the model predictions, we visualized the consistency in model predictions over the course of the robot's observations. This enabled us to devise a *fluctuation metric* that we used to select a model for our study. Using the prediction visualization, in conjunction with the feature visualizers, we found two potential culprits for the oscillating predictions. First, one of our object detectors had a very high false positive rate on study-related building blocks; we switched off this detector. Second, we recognized the value of capturing human pose for interruptibility classification; therefore we added a pose estimator to our perception pipeline.

In training, testing and visualizing our model, we were greatly aided by the fact that we had implemented a processing pipeline that could automatically derive new features from old data and reassociate ground truth labels with them.

3.3 Insights

Model evaluation and selection was one of the most important pieces for our user study. Unfortunately, the evaluation metrics commonly used in machine learning (e.g., F1 score) did not prove sufficient in providing accurate evaluations of the model by themselves. Instead, visualizations of the model and data, in conjunction with an ensemble of evaluation metrics (F1 score and the fluctuation metric) helped us to (1) adequately redesign the study, (2) improve our classification system, and (3) pick the best model for our study.

4 LESSON 2: PARTICIPANT BEHAVIOR

We needed to employ deception in our study to avoid priming participants into exaggerating their behaviour to showcase interruptibility. Unfortunately, the deceptions left participants with room to interpret the rules and boundaries of the study for themselves, leading to undesirable and even adversarial participant behavior.

[1] Participants could hear this response.

[2] For example, ignoring a build on the main tablet, or picking up the robot tablet and then replacing it without completing the robot build

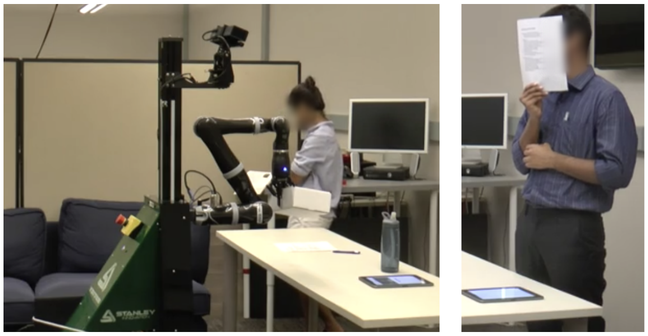

Figure 7: Examples of unexpected participant behavior. (Left) Apathetic participant ignoring main task and robot task. (Right) Adversarial participant hiding from the robot

4.1 Problems

There were two broad, not mutually exclusive, classes of problematic participant behavior. First, participants attempted to deceive the robot in its fictitious task, sometimes attempting to go so far as to hide themselves from it (Fig. 7 (Right)). Second, participants were sometimes apathetic (Fig. 7 (Left)). The problems were exacerbated in our first iteration of the study where we had two participants work side-by-side with the goal of having our robot interrupt the most interruptible participant: in a notable study trial, coquettish participants refused to interact with the robot *and* ignored the instructions provided to them.

4.2 Solutions

Our experience with adversarial behavior in the first iteration of the study engendered the following solutions:

(1) We changed our hypotheses and metrics so that we could show effective robot behaviour with a single participant. Recruitment became easier and it eliminated coquetry.
(2) Distractor items were introduced to occupy participants in an acceptable manner when they were idle.
(3) We improved upon our narrative. The simulated breaks were a key addition that allowed experimenters to interact with participants during the course of the study.
(4) We improved our mechanisms for participant monitoring with a webcam for observing the build area, and a method of observing the participants' behavior on the tablets.
(5) We questioned participants on their behavior post-study.
(6) We established a data sanitation protocol to salvage study data via the external camera.

Finally, we updated our study protocol to include a criterion for stopping a study session early in an effort to not waste time.

4.3 Insights

Our original study design was based on assumptions that people would work diligently, respond to the robot in the absence of other tasks, and most importantly, would follow instructions even without the presence of experimenters. Once those assumptions were violated, we were able to design a study that prepared for apathetic or adversarial participants and still yielded valuable data for our research questions. In fact, after some trials in the final design, post-study interviews informed our quantitative data analyses by prompting a search for evidence of self-reported effects. Finally, we found our early-stopping criterion to be very valuable.

5 LESSON 3: UNEXPECTED CONFOUNDS

In choosing a skill-based task to measure the effects of interruptions on task throughput, we were aware of the potential for participant building skills to serve as a large confound in our data. However, we had no method to ground our expectation on the skill confound. We constructed multiple builds of varying difficulties and designed flexible timing conditions in the first iteration of our study to allow differing levels of build skill to dictate build times. We expected multiple participants of various skills to later control for the confound in the data analysis.

5.1 Problems

An analysis of the data from the first iteration of the study revealed that our control of the confounding variable of build skill had been inadequate. The variance in our metrics due to skill level promised to dwarf the variance in our metrics as a result of the study condition.

5.2 Solutions

In the second iteration of the study, we imposed a stringent time schedule on all participants. In addition, we greatly simplified the tasks provided to the participants; assigning them only the simplest builds. Pilot studies allowed us to verify that these changes led to a more noticeable effect of the study condition on our metrics.

5.3 Insights

Although we were aware of the potential confound of participant build skill, we were unable to adequately predict its effect on our data. We were also lazy in designing around potential confounds, instead hoping to be able to control for skill in the data analysis. Our data from the first iteration of the study would have been insufficient for answering our research questions, even if we hadn't run into problems of model evaluation and undesirable participant behavior. Explicitly controlling for the confound of skill greatly improved quality of our data, and allowed us to draw meaningful conclusions to our research questions.

6 SUMMARY AND CONCLUSIONS

In order to completely avoid problems in a user study, we advise removing both the robot and the participants from the study design. Unfortunately, it is impossible to design a user study without both. In this work, we reviewed three main lessons we learned from a recent user study involving deceptions featuring an autonomous mobile robot and confounding human participants. We examined the importance of model evaluation and selection, proper participant motivation and instruction, and active control of confounding factors. While it is impossible to completely solve the problems that we encountered, we hope that our advice on how to identify, approach, and solve these issues can prove useful to the research community in HRI.

REFERENCES

[1] Siddhartha Banerjee and Sonia Chernova. 2017. Temporal Models for Robot Classification of Human Interruptibility. In *AAMAS*. 1350–1359.