INVITED PAPER

# Building and Verifying a Predictive Model of Interruption Resumption

*Help from a robot, to allow a human storyteller to continue after an interruption, is explored; results indicate a bright future for effective human–robot interaction.*

By J. Gregory Trafton, Allison Jacobs, and Anthony M. Harrison

**ABSTRACT** | We built and evaluated a predictive model for resuming after an interruption. Two different experiments were run. The first experiment showed that people used a transactive memory process, relying on another person to keep track of where they were after being interrupted while retelling a story. A memory for goals model was built using the ACT-R/E cognitive architecture that matched the cognitive and behavioral aspects of the experiment. In a second experiment, the memory for goals model was put on an embodied robot that listened to a story being told. When the human storyteller attempted to resume the story after an interruption, the robot used the memory for goals model to determine if the person had forgotten the last thing that was said. If the model predicted that the person was having trouble remembering the last thing said, the robot offered a suggestion on where to resume. Signal detection analyses showed that the model accurately predicted when the person needed help.

**KEYWORDS** | Cognitive robotics; cognitive science; human–robot interaction; interruptions and resumptions

Manuscript received October 29, 2010; revised August 30, 2011; accepted October 12, 2011. Date of publication January 13, 2012; date of current version February 17, 2012. This work was supported in part by the U.S. Office of Naval Research under funding documents N0001411WX20407, N0001409WX20173, and N0001411WX20474 to J. G. Trafton. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Navy.
The authors are with the Naval Research Laboratory, Washington, DC 20375-5337 USA (e-mail: greg.trafton@nrl.navy.mil; jacobs.allie@gmail.com; anthony.harrison@nrl.navy.mil).

Digital Object Identifier: 10.1109/JPROC.2011.2175149

## I. INTRODUCTION

As computers and machines become more intelligent, they will need to deal more (not less) with people. As long as people are "in the loop" autonomous systems will need to interact with them, help them solve problems, keep them on task, remind them of missed appointments, etc. If an autonomous system can predict what a person needs and when they will need it, that system will have better autonomy and be a better system overall than a system that cannot predict what a person will do. Unfortunately, the vast majority of autonomous systems today are barely able to interact with people in a simple manner, much less predict what they are thinking and act upon it. Our primary goal in this paper is to show our approach to building predictive models of human behavior, and how we use computational cognitive models to improve human–robot interaction.

The context we are focusing on is resuming after being interrupted. With the rapid rise of communication technologies that keep people accessible at all times, issues of interruptions and multitasking have become mainstream concerns. For example, *Time* magazine [1] and the *New York Times* [2] both reported stories about interruptions and multitasking and how they affect performance. The information technology research firm Basex issued a report on the economic impact of interruptions, which they estimated to be around $588 billion a year [3]. Given the prevalence of interruptions, building systems that can help remind an individual what they were doing or where they were in a task can have a large impact on individual and group productivity.

Being interrupted also greatly increases the number of errors [4]. People will frequently repeat a step that has already been performed or skip a step that needs to be

 

performed after an interruption. Sometimes these errors are irritating (e.g., destroying a meal by leaving out a crucial ingredient), but sometimes they can have disastrous consequences (e.g., taking medicine twice or not configuring the flaps for airplane takeoff). The research described here is applicable to these domains, but this report will focus on a common, everyday task: being interrupted while telling someone a story or giving instructions. This information-passing task is an excellent domain for studying the interruption/resumption process for several reasons. First, because it is so common to get interrupted while talking to a friend, it is easy to collect data. Second, providing ordered information to another person is a general class of problems that include recipes, checklists, story telling, direction giving, etc.

For example, in the middle of giving you instructions on how to operate a new device, your friend needs to take an important phone call for a few minutes. When she comes back to tell you the rest of the instructions, what does she do? If she cannot remember exactly where she left off, you may remind her or she may resume where she thought she left off (which may or may not be correct). If your friend was telling you a story, she may simply start somewhere close to where she left off. For the remainder of the paper, we will focus on building a process model of exactly what the interlocutor is doing as she attempts to resume the conversation, then using that process model to allow a robot listener to facilitate the interaction.

We will use two theoretical frameworks: transactive memory [5] and the memory for goals (MFG) theoretical framework that we have previously used to interpret how subgoals are suspended and resumed in a problem-solving task [6], [7] and which has since been used to interpret the time costs of interruption [6]–[8], factors affecting post-completion error [9], and sequence errors [4].

### A. Transactive Memory

Transactive memory occurs when two or more people work together on a common task. Each group member becomes responsible for remembering certain items, usually based on past experience. For example, spouses spontaneously divide up things to remember based on past history, areas of expertise, and context. The family gardener, for example, will not only be able to remember the names of different plants, but other members of the group will expect the gardener to remember those plants (and whether they need to be removed), and so will not commit resources to remember that information. The overall effect here is that the group as a whole can systematically remember more than any individual.

### B. The Memory for Goals Model

MFG is a theory about how people remember and retrieve goals. It has been instantiated in the cognitive architecture adaptive control of thought—rational/embodied

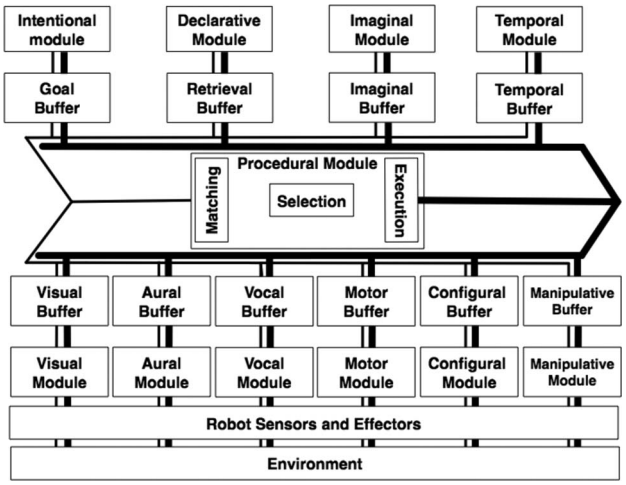


**Fig. 1.** ***Schematic of ACT-R/E.***

(ACT-R/E) [10]–[13] that is shown in Fig. 1. ACT-R/E will be described first to provide some context for MFG.

ACT-R/E is an embodied version of ACT-R [10], [11]. It consists of a set of modules, each specialized to process a different type of information. For example, the goal module keeps track of current intentions, the declarative module retrieves information from memory, the imaginal module keeps track of intermediate products, the temporal module estimates how much time has passed, the visual module identifies objects in the visual field, the aural buffer hears sounds in the auditory field, the vocal module speaks, the motor module moves the body, and the configural and manipulative modules perform spatial processing [14]–[16]. The procedural module is a production system which responds to what is in the buffers at any given time step.

ACT-R assumes a mixture of parallel and serial processing. Each module operates in parallel: the visual system is processing the entire visual field; the declarative memory system is searching for a specific memory in parallel, etc. However, there can only be a single object in each buffer at one time. The basic cycle consists of the contents of all the buffers being matched against the rules stored in procedural memory. A single rule is then chosen on the basis of its utility (the rule that has the lowest expected cost while having the highest expected probability of succeeding), and this rule carries out a set of actions, which it communicates to the other modules through their respective buffers. More detailed information about ACT-R is available elsewhere [10], [11], [17]. The MFG model is focused on the declarative module, so that module will be described in some detail.

All memories in MFG have an activation associated with them. The MFG model inherits two basic processing assumptions from ACT-R. The first is that when central cognition queries memory, memory returns the relevant

item that is most active at that instant. The second is that the activation of a given memory element fluctuates noisily from moment to moment about a mean value. Activation of a memory element has three components: history, context, and noise. The history of a memory element is defined in [7]

$$m = \ln\left(\frac{n}{T^{-d}}\right) + \varepsilon. \quad (1)$$

$m$ is the activation of the memory element, $n$ is how often the memory has been sampled over its lifetime, $T$ is the length of the goal's lifetime from encoding to present, and $d$ is the decay rate. $\varepsilon$ is activation noise, which governs the variance of the zero-mean logistic noise distribution sampled for the activation of each control code on each system cycle. Thus, as a memory is sampled or strengthened or rehearsed, it gains in activation. As time passes, it loses activation. In order to remember a memory element that does not have the highest absolute activation (e.g., a very recent memory), priming through context (instantiated through simple co-occurrence of cues) associatively links different memory elements and boosts an element's activation [7]. In the current model, these memory elements are episodic codes—a memory of an individual event that can then be remembered later. To retrieve a memory element, the system makes a request (e.g., what did I have for breakfast this morning?) and the most active element matching those specifications is retrieved.

Episodic codes serve the place-keeping function of interest as the storyteller tells the story. These episodic codes have been used in models of task switching [18] and well-known procedural tasks [4]. While executing a well-known task (e.g., making coffee or tea), the model posits that people create an episodic code each time a step is executed. For example, separate episodic codes are created as a person pours hot water into a cup, puts a teabag into the cup, and adds sugar to their tea.

Here, we assume that as a story is retold, the system consults its declarative knowledge, which we assume is well learned (e.g., the story is known to the teller and not being made up on the fly), uses that knowledge to create an episodic memory of the retelling, and uses that episodic memory to guide its telling of the story. This episodic code essentially represents the information, "this is where I am in the story," and serves as a reference point for any component process that may have to run in the course of that step. While in theory the episodic code could code any level of granularity, we assume that people encode the gist of the story [19]–[21], not the low level features (e.g., the specific words), so episodic codes in this context contain gist information. Because these episodic codes decay, the current one will always be the most active (modulo effects of activation noise), so any process can reliably assume that whatever code it retrieves is where in the story-retelling process it is.

After an interruption, the system can use an episodic code to regain its place in the story. Here we assume that the interruption most often occurs between gist components—that is, after one part of the story has completed and before the next has begun. This assumption is incorporated in the model, such that, after an interruption, the model assumes that the episodic code it retrieves was for the most recently completed gist.

Sometimes, of course, people will continue a story at a place that is not exactly where they left off. The MFG model (see Fig. 2) makes a specific prediction about the pattern of results that should occur. After an interruption,

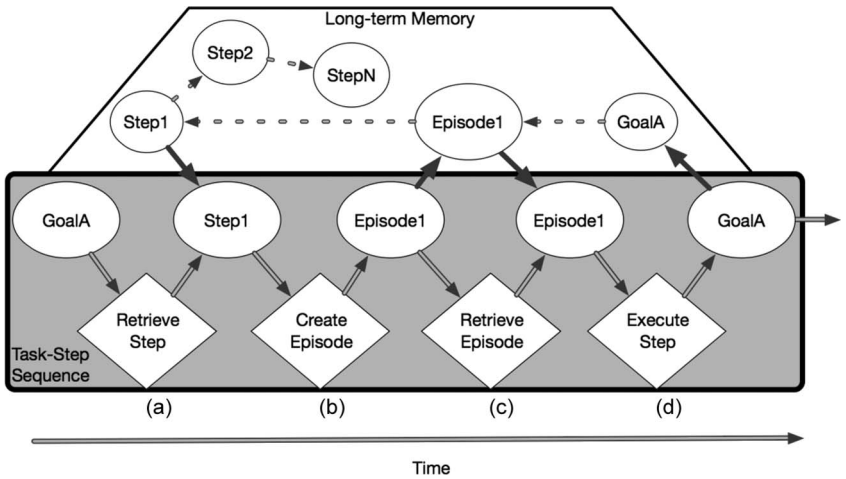


**Fig. 2.** *Schematic of the MFG model. For each discrete goal step (retelling a gist element from the story), the model retrieves the next gist element (a), creates and encodes an episode for that gist (b), immediately retrieves that episode (c), then executes the step by telling the gist (d). Step repeats are due to incorrect retrievals at (c), whereas skips are due to interruptions at (d).*

the system retrieves an old episodic code and assumes that it represents the most recent gist. Most of the time, this assumption will be correct. Sometimes, however, an older episodic code will intrude, even though it is more decayed than the most recent old code, due to noise in activation levels. In this case, the model will repeat something that it has already said. For example, if the second-most recent episodic code is retrieved, the storyteller will tell the last two gist events over again to the listener. From a social or functional perspective, this is probably beneficial because the listener will be given context for where the story is before continuing on to the parts that had not been heard before. Thus, the MFG model predicts that the storyteller should repeat the last thing that was said the majority of the time, occasionally repeat previous gist items, and only rarely actually skip to a part of the story that had not been told yet.

Occasionally the episodic code will have decayed so much that it will be very difficult to retrieve. In extreme cases (e.g., an interruption that lasted three days), the episodic code will have decayed so much that it is extremely difficult or impossible to retrieve. The storyteller in these cases may just start over. In less extreme cases, the storyteller may need to try different internal cues or work up to where they were in order to continue the story.

## II. EXPERIMENT 1

Our primary goal with this experiment was to explore how people resumed telling a story after being interrupted and to determine where they resumed after being interrupted. A confederate listener was used so that we could make sure that the listener would behave in the same manner across conditions and participants. We also were interested in how long it would take for people to resume.

### A. Method

*Participants:* Forty three undergraduate George Mason University students participated for course credit. All participants were female to match the gender of the confederate listener. Four participants asked to have their data destroyed after the experiment, leaving a total of 39 participants. The average age of participants was 19.7 years old.

*Task and Materials:* Each participant was asked to read three total pages of a soap-opera-like story. The first two pages (six paragraphs, 1164 words) were the primary story. The last page (three paragraphs, 494 words) was the interrupting task. The participant's task was to read the story, then retell the story.

*Design and Procedure:* The design of this experiment was a between-participants 2 (person, video camera) × 2 (interrupted, control). Participants in the control condition were run to verify that the location of the interruption was not an especially difficult part of the story. Only 13 participants were run in the control conditions; everyone else was in the interruption condition.

Participants in the person condition were introduced to another "participant" (actually a confederate) and were given either two (in the interruption condition) or three (in the control condition) pages of a story to read. After they had finished reading the story, they retold the story to the confederate.

After retelling approximately two-thirds of the primary story, participants in the interruption condition were interrupted by the experimenter at a predetermined location. The experimenter told these participants that the third page of the story had been accidentally separated from the first two pages. She asked them to stop retelling, read the third page, and then to resume telling the primary story from the point where they had been interrupted. During the interruption, the confederate quietly sat and waited until the participant had finished reading the third page. When the participant retold the story, the confederate did not help in any way, even if asked. Participants in the control condition were given all three pages to read and were not interrupted. In both conditions, the confederate nodded and showed interest throughout the entire story retelling.

Participants in the video-camera condition were told that they were telling the story to a video camera, and that the film would be shown to another participant at a later date. Half the participants were interrupted and half were not interrupted at all.

All participants were videotaped. After debriefing, they could choose to have their videotape destroyed. Four participants (one from each condition) asked for their videotapes to be destroyed, which occurred immediately.

*Measures and Coding:* Resumption lag (RL) was coded as the time from the end of the interruption (or the intended point of interruption in the control conditions) until the participants began to fluently resume the primary story. Interruption duration was coded for those participants in the interruption conditions as the time between the start of the interruption and the end of the interruption.

Participants in the interruption condition had their videotapes transcribed right before the interruption and right after the interruption. Participants in the control condition had their videotapes transcribed right before the interruption location through the next gist item that they described. Three utterances before and three utterances after the interruption location were transcribed for all participants.

Participants' utterances were classified by whether they appeared to need help at the resumption point and where they actually resumed when they did resume.

Participants were coded as needing help if they asked for help (either to the confederate or to themselves), stating that they did not remember where they were, or using

a large number of speech disfluencies or fillers (e.g., "soooooo," "uhhhh," "ummmm," etc.) that were not present before the interruption. All participants were eventually able to resume the story at some point, yet not always at the correct resumption point.

To code the location of the resumption, we coded the gist of the story around the interruption location, and marked it as either "repeat" (e.g., a gist utterance that was a repeat of what had already been said), "correct" (e.g., the next gist that occurred in the story) after the last thing that was uttered), or "skip" (e.g., an utterance that skipped the correct resumption gist).

### B. Results

The empirical results were analyzed using an analysis of variance (ANOVA). In the following analyses, the $F$ statistic is used for testing the differences between groups (with degrees of freedom in parentheses), the numerical value of the statistic, a report of the mean square error (MSE; a measure of variability), and the probability of the result occurring by chance [either less than 0.05 or not significant (n.s.)]. A post-hoc analysis using the Holm test was used to highlight where any differences occur if there are more than two groups.

When the data were frequency based, a chi-square analysis was used, which deals with nonparametric data more robustly than an ANOVA.

*Interruption Duration and Resumption Lag:* Participants who got interrupted spent approximately the same amount of time reading the interrupted story ($M = 229.5$ s), whether they were telling the story to a person ($M = 242.7$ s), or telling the story to a video camera [$M = 222.6$, $F(1, 24) = 0.6$, MSE $= 8351$, *n.s.*]. Not surprisingly and consistent with previous research, participants who got interrupted took longer to resume than participants who did not get interrupted [$F(1, 35) = 11.9$, MSE $= 81.9$, $p < 0.05$]. As Table 1 suggests, however, neither the effect of listener [$F(1, 35) = 0.7$, MSE $= 81.9$, *n.s.*] nor the interaction between listener and interruption [$F(1, 35) = 0.1$, MSE $= 81.9$, *n.s.*] approached significance. In terms of response time, there was a strong impact of the interruption, but no effect at all of who the participant was telling the story to.

*Needing Help:* To establish inter-rater reliability (IRR), one coder coded all the data for disfluencies and decided whether that individual needed help. A second coder then coded 67% of the participants, also making a decision on whether each individual needed help. The two coders agreed 88% of the time, $\kappa = 0.76$, $z = 3.9$, $p < 0.01$. A kappa of 0.76 is considered extremely good. Disagreements were resolved through discussion. This coding showed that the coders could reliably agree when a participant needed help (or would have appreciated someone reminding them what they had said last). Table 2 shows two examples of people who seemed to need help and two examples of people who were able to resume without problems.

Table 1 Means and Standard Deviation (in Parentheses) of the Resumption Lag for All Four Conditions. All Measures Are in Seconds

| | Human | Video |
|---|---|---|
| **Interrupted** | 13.2 (11.1) | 11.5 (10.8) |
| **Control** | 1.1 (.8) | 1.5 (.6) |

Table 2 Example Utterances of Participants Needing Help and Not Needing Help

| Utterance | Coding |
|---|---|
| I don't remember where I left off at, um. It'll come back to me. Do you remember? Hold on. Um, I don't know, but I'll just go back, maybe it might be a little further than that, I mean before. Um, I think back at the parents'. So the dad Adam, um, no ok, so | Needed help |
| Ok, um, ok, ss- ok so, um, soooooo, | Needed help |
| OK, so, Laura came over… | Did not need help |
| Alright, so Haley's Dad tried to … | Did not need help |

All participants who were not interrupted were able to tell the story smoothly and without disfluencies, suggesting that the location chosen for the interruption was not an especially difficult part of the story. Of the participants who did get interrupted, 77% of the participants who told the story to a person asked for or needed help, while only 35% of the participants who told the story to a video camera acted like they needed help; this difference was statistically different, $\chi^2(1, N = 26) = 4.2$, $p < 0.05$. As Fig. 3 suggests, participants who were not interrupted were fluent at retelling the story at the location that participants in the interruption condition were interrupted, but the interruption caused an increase in RL, omnibus ANOVA $F(2, 36) = 21.4$, MSE $= 49.7$, $p < 0.001$. Specifically, participants who did not need help were slower at resuming the story than participants in the control condition (Holm adjusted $p < 0.01$) and faster than participants who did need help (Holm adjusted $p < 0.01$). Participants who did need help were also much slower than participants in the control condition (Holm adjusted $p < 0.01$).

*Where People Resumed:* After participants resumed, where in the story did they resume? As Fig. 4 suggests, participants did not differ in their resumption patterns if they needed help or not [$\chi^2(2, N = 26) = 0.25$, $p = 0.88$], nor were there any differences in whether the participant told the story to a video camera or a person [$\chi^2(2, N = 26) = 1.2$, $p = 0.56$]. Participants did, however, resume more frequently by repeating what they had already said than resuming at the "correct" location or

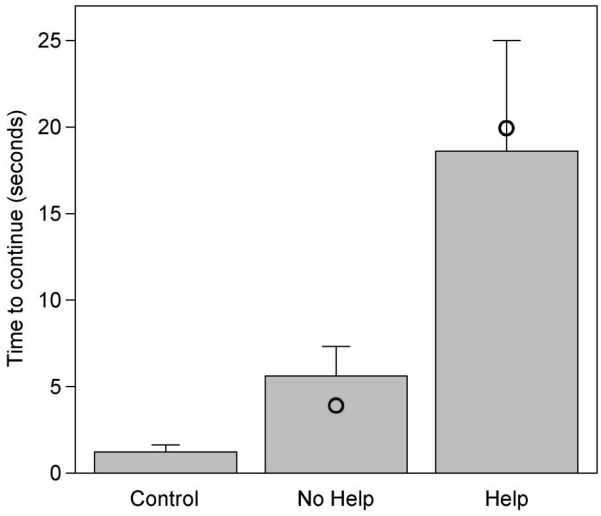


**Fig. 3.** *The amount of time it took people to resume (bars) and ACT-R/E model fits (circles). Error bars are 95% confidence intervals.*

skipping one or more story gist statements, omnibus $\chi^2(2, N = 26) = 19$, $p < 0.05$; Holm adjusted $ps < 0.05$. This pattern of results is consistent with the MFG account that people will attempt to remember the last episodic code they talked about and resume from there.

### C. Discussion

In summary, people had no problem retelling a story they had just read when there was no interruption. When there was an interruption, however, people took longer to resume than when there was not an interruption, though it did not matter if the person was telling the story to a physically present listener or simply to a video camera. If the story was being retold to a physically present person, the storyteller did want help from the listener more than 75% of the time. Interestingly, when people needed help, they seemed to have a much harder time remembering where they should resume as shown by the fact that it took them more than three times as long to be able to continue the storytelling as people who did not need help. When people did resume, they typically repeated part of the story they had already described.

Consistent with a transactive memory approach, people relied on their partner when there was a partner available. The speaker evidently assumed that the listener would be able to help. This strategy was unsuccessful because the confederate was instructed not to help the participant, but the result certainly shows the willingness of the storyteller to use a different memory source. When there was no one to rely on (e.g., when they were telling the story to a video camera), the storyteller needed help much less often. This result suggests that when there is no one to help them remember the last thing they said, people will use their own memory processes rather than rely on someone else's.

These results are also broadly consistent with an MFG approach. Generally, people create an episodic trace as they retell a story, and after an interruption they attempt to remember where they were by retrieving this episodic trace. The MFG model's prediction that when people resume, they will usually repeat the last thing they said was confirmed in this data set. To capture the details of this resumption process, an MFG model was created.

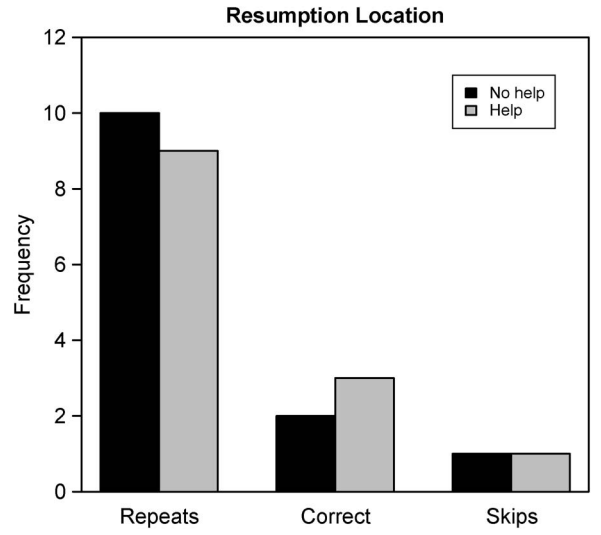


**Fig. 4.** *Where participants resumed.*

## III. MODEL DESCRIPTION

We used the MFG framework described earlier to develop a cognitive simulation of the storytelling task.

The model has a moderately lean representation of the processing that occurs during the storytelling task. Human participants must iteratively recall the gist of the next story element from memory and then elaborate upon that element with any relevant details. This process is repeated until the story is complete or the individual is interrupted. The elaboration process is primarily one of natural language generation and not the focus of this work. Instead we focus on the process of creating, storing, and recalling each gist element with a focus on resumption after an interruption.

MFG postulates that at each discrete step in the execution of a task, an episodic control code is created [Fig. 2(b)]. This episodic *tag* effectively marks the position in the task by virtue of its existence in declarative memory. The episodic tags effectively create an associatively linked list of markers to completed steps in a task. If an interruption occurs, the episodic tag can be used to return to that point in the task. At resumption time [Fig. 2(c)], the model attempts to retrieve the most active episodic tag for that particular gist. If successful, it will use that tag and the associated gist element to retrieve the next to-be-reported

gist element, allowing the model to continue the task [Fig. 2(d)]. The model does not check to see if the episodic tag is, in fact, correct. If the model fails to retrieve the relevant episodic control code, one of two options is available to the model. If there is a listener available, the model will merely ask for help. If, however, there is no listener available, the model will try again to remember an episodic tag, and will make repeated attempts until it successfully retrieves an episodic tag. This model provides a process description of transactive memory: the listener can serve as another memory source, but is only used if the person cannot remember themselves.

As described earlier, the model depends critically upon the basic properties of declarative memories. When a retrieval is requested, the matching chunk with the highest activation is returned [Fig. 2(a) and (c)]. The activation of a chunk depends upon the recency and frequency of that chunk's use [see (1)], in addition to a contextual element (i.e., priming) and stochastic noise. Immediately after encoding, episodic tags have very high activations, but they decay quite quickly after that. Retrieval during uninterrupted performance [Fig. 2(c)] is facilitated by the contextual priming through activation spreading from the current focus of attention (the priming constraint from [7]). As a tag's activation decays, other tags may, temporarily, outrank it as noise effects become more significant. This error gives rise to the repeats seen in the storytelling data. Single-step skip errors seen in the data arise due to when the interruption occurs. If the interruption occurs after episodic tag encoding, but before communicating the gist, the correct tag may be retrieved even though the step was not actually completed [Fig. 2(d)]. Not surprisingly, this predicts that the chance of a failing to retrieve the correct episodic control code increases with the duration of the interruption.

The model currently does not account for the reading and encoding of the story. Rather, we assume that the gist of the story is perfectly recalled and that the errors observed are due to episodic failures, not knowledge gaps. Since the interruption in the behavioral study was to read subsequent story sections, the model simply waits for the average amount of time subjects spent reading the last section (approximately 230 s). Because the duration of the interruption varied from subject to subject, the model makes use of ACT-R's temporal module [22]. The temporal module provides some variability in time estimation, allowing some robustness to the interruption duration measure. In this context, we assume that the interrupting task itself does not directly impact the error rate; it is the time spent away from the primary task that increases the probability of making an error.

Several model parameters interact to affect the behavior described above, all of them affecting activation dynamics. Activation noise ($\varepsilon$) in (1) is set to 0.25. Increasing the activation noise allows prior episodic tags to intrude more frequently. A second parameter is the decay rate [$d$ in (1), set to the default value 0.5]. The decay rate controls the speed with which activations deteriorate with time. The faster the decay, the more sensitive the model will be to the duration of the interruption. A third parameter is priming (set to 1, described in more detail in [7]) which provides contextual priming to help retrieve the relevant control code. Strengthening the priming makes the retrieval of episodic tags more robust, by strengthening the immediate context's (i.e., the goal) effect on the episodic tag's activation. Finally, the retrieval threshold for declarative memory (set to $-1.1$ from the default value of 0) allows ACT-R to devote more time to retrieving a memory than it normally would. These parameters are well within the range of values usually used within the ACT-R community. Sensitivity analyses have shown that the qualitative data pattern is stable over a large parameter space.

To reproduce the empirical data, we ran 2000 simulated trials with a (virtual) listener available and 2000 simulated trials with no listener available. The primary measures of interest were how long the model spent on the interruption, how often the model asked for help when a person was available, and how long it took to resume both when a listener was available and when a listener was not available.

### A. Model Fit

Recall that participants took an average of 230 s to read the last page of the story. We set the temporal module to wait for an average of 227 s. This was an important aspect of the model's success because the model's ability to retrieve an episodic code depends critically on the amount of time spent on the interruption.

Also recall that, when there was a listener available, the participant asked for help or acted like they wanted some help 77% of the time. When the model attempted to retrieve an episodic code after the interruption, it was unable to do so 80% of the time. After this first failure, since there was a listener available, the model asked for help. When the model did retrieve an episode, the episode was correct or a simple repeat the vast majority of the time.

Probably the most important measure, however, is the RL after the interruption for both conditions. As is evident in Fig. 3, the model matches the data quite well; root mean square deviation (RMSD) = 1.5. Critically, all model data are within 95% confidence intervals of the empirical data.

## IV. SUMMARY OF EXPERIMENT 1

The MFG model showed an excellent fit to the experimental data. However, as with all model fitting paradigms, it is possible that the model will not generalize to other situations because the cognitive processes, parameters, participants, or experimental task may be idiosyncratic or the model was overfit. The approach we have chosen to deal with these issues here is one of strong cross validation

and prediction. We will take our current model and run it on our robot as it plays the part of a listener. The model will attempt to predict when someone needs help and then provide a reminder to the storyteller. Participants will be from a different group than the first experiment as well. If the overall system is able to successfully help people in their resumption process, we will assume that the model accurately describes the cognitive processes involved in interruption resumption.

## V. EXPERIMENT 2

Our primary goal with this experiment was to determine if our MFG process model of resumption after an interruption could predict when people would need help. Our MFG model was put on our robot platform (described below) and run as a model of the speaker. After an interruption, if the model was not able to retrieve where the person left off, the model assumed that the person had forgotten as well, so would spontaneously attempt to help. If the model was able to retrieve where the person left off, the model assumed that the person had as well and would not offer help.

### A. Method

*Participants:* Twenty two employees at the U.S. Naval Research Laboratory (NRL) participated in this study. Eleven of the participants were men and 11 of the participants were women. The average age of participants was 42 years old.

*Task and Materials:* The story and story-retelling task were identical to experiment 1.

*Robot Description:* Our current robot platform is the mobile–dexterous–social (MDS) robot [23]. The MDS robot neck has 18 degrees of freedom for the head, neck, and eyes allowing the robot to look at various locations in 3-D space. Perceptual inputs include two color video cameras and an SR3000 camera to provide depth information. Fig. 5 shows a photo of Octavia, the MDS robot used in this study.

Our usual operation is to use ACT-R as our robot controller (see [12], [16], and [24]–[26] for our approach on this). In this case, for the responsive robot partner (see below), we utilized two ACT-R models. The host model directed the robot's listening behavior and interaction with the participant. The host model was also responsible for the execution of the slave MFG model of the participant. This slave model of the participant starts at the interruption point and runs while the participant reads the subsequent story elements. When the participant was done reading, the experimenter signaled the slave MFG model that the resumption had begun. At this time, the MFG model attempts to recover its previous episodic tag. The model's success or failure is reported back to the host model. If the MFG model fails to retrieve its previous goal, the host prompts the participant; otherwise, the host resumes its listening behavior.

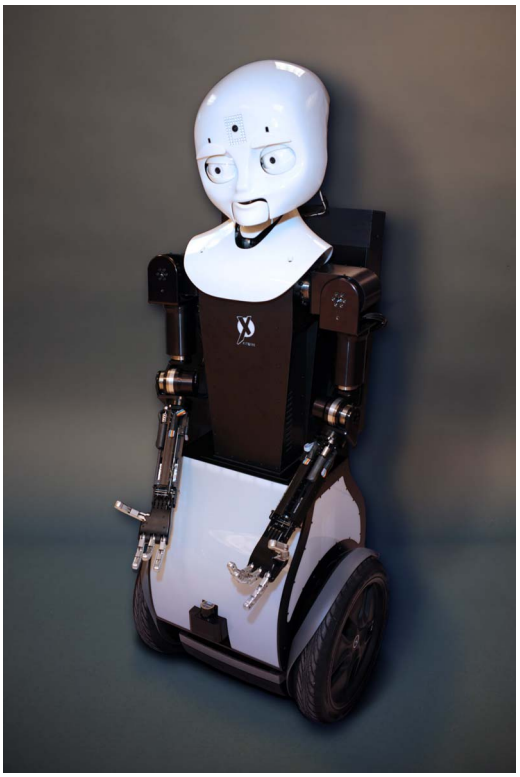

**Fig. 5.** *An image of the MDS robot that was used in experiment 2.*

The responsive host model's listening behavior consisted of simple behavioral components. The robot visually fixated on and tracked the speaker using a fiducial [27] attached to a hat worn by participants. Additionally, the robot would occasionally blink and nod during its interaction with the participant. This small level of interactivity was used to push people towards believing that the robot had basic social competencies [28]. People will, for example, follow a robot's gaze [29], attribute personality and gender stereoptypes to computers/robots [28], [30], and willingly anthropomorphize robots with very little evidence that the robot can think or act for itself [31]–[33]. We hoped that the combination of following the person as they moved their head around, nodding, and blinking would encourage the participant to feel like the robot was actively listening to the story.

*Design and Procedure:* The design of this experiment had two conditions: a responsive robot partner and a control condition with an unresponsive robot partner. All participants in this study were interrupted. As in experiment 1, the experimenter "forgot" to give the participants part of the story; after they had finished reading it, they were asked to resume where they left off.

In the control condition, the robot did not move nor did it help the participant remember where they left off; this condition was similar to our video camera condition from experiment 1. In the responsive robot condition, the robot not only nodded and followed the person's face as she/he told the story, but if the MFG model could not remember where the person left off, it assumed that the person could not either, so provided the participant with an appropriate prompt (e.g., "I think you were telling me about Haley's father").

This specific prompt worked because the experimenter interrupted the user at the same point in the story every time, so the resumption utterance made sense for participants. If a functional natural language system was able to provide us gist information, we could have used that, but at this point in time, natural language systems are not sufficiently advanced to provide this information to our models. If the person was able to resume the story before the model had made a decision, the experimenter cancelled the MFG model. In either case, the robot resumed its listening behavior.

Note that the MFG model partially tailored itself to each participant. If the participant was a fast reader, the interruption duration would be shorter than average and the probability that the person (and the model) would remember where they left off was higher than average. In contrast, if a participant was a slow reader, the interruption duration was longer, and the episodic trace would have had a longer chance to decay, so would be more difficult to remember. Neither the host model nor the MFG model used any social cues of the user (e.g., long pauses or disfluencies, facial expressions, asking for help, etc.). Thus, this was a very pure predictive MFG model.

After the experiment was completed, all participants were given a short exit questionnaire and debriefed.

As in experiment 1, all participants were videotaped. After debriefing, they could choose to have their videotape destroyed. No participants in experiment 2 asked for their videotapes to be destroyed.

*Measures and Coding:* Interruption duration was again coded as the time between the start of the interruption and the end of the interruption. Participants' videotapes were again evaluated for whether they needed help using the same criteria as in experiment 1. In the responsive robot condition, we also recorded whether the robot offered help to the participant. As in experiment 1, all participants were eventually able to resume the story.

The exit questionnaire consisted of three questions. 1) "Please rate how *natural* the robot was as a conversational partner." 2) "Please rate how *useful* the robot was as a conversational partner." 3) "Please rate how *comfortable* you were with conversing with the robot." Participants answered each question on a 1–7 Likert scale where 1 was "completely unnatural/nonuseful/uncomfortable" and 7 was "completely natural/useful/comfortable."

## B. Results

*Interruption Duration:* Participants spent an average of 251 s reading the interrupted story. Their interruption durations ranged from 128 to 709 s. Participants in the responsive condition ($M = 260$ s) did not spend any longer reading the interrupted story than participants in the control condition ($M = 233$ s), $F(1, 20) < 1$, MSE $= 17206$, *n.s.*

*Needing Help:* Participants did not differ in the amount of help they needed across conditions (79% versus 87%), $\chi^2(1, N = 22) = 1.1$, $p > 0.10$.

*Robot Helping Evaluation:* In the control condition, the unresponsive robot did not offer help at all (by design). In the responsive robot condition, the robot helped 80% of the time. Clearly, the responsive robot attempted to help participants when it thought they needed help. To evaluate whether the robot helped people when they actually needed help, we performed a signal detection analysis.

In order to quantitatively determine the overall robustness of the model in its helping behavior (i.e., to determine whether the robot helped people when they actually needed help), a signal detection/$d'$ analysis was performed [34]. To perform this analysis, we calculated the hit rate and the false alarm rate. The hit rate is the number of times that the person needed help and got help divided by the total number of times that the person needed help (regardless of whether the robot helped). The false-alarm rate is the number of times the robot helped but the person did not need help divided by the total number of times the person did not need help. A $d'$ score is the $z$ score difference between the hit rate and the false alarm rate; a higher $d'$ is better than a lower $d'$.

In the responsive-robot condition, the hit rate was 1 while the false alarm rate was 0.3, leading to a $d'$ of 4.8, suggesting that the robot was helping primarily when the person needed help and not helping when the person did not need help.

In the control condition, the robot did not help, of course. However, it was possible to run the model multiple times and to match the model inputs for each participant in the control condition (in this case, interruption duration). Because each model run is probabilistic, we averaged each of ten model runs to give us a probability of helping. If the probability was 70% or greater, we assumed that the model would have helped the participant if the predictive model had been running. We found that in the control condition, the hit rate was 0.71 while the false alarm rate was 0, leading to a $d'$ of 4.8. This result suggests that if the memory for goals model had been activated in the control condition, it would have helped a majority of participants.

*Subjective Evaluation of Robot's Helping Behavior:* Another way to evaluate the performance of the robot was to determine how participants felt towards the robot—their

subjective impressions. In general, participants were quite comfortable with the robot ($M = 5.1$ on a seven-point scale). Participants who received help ($M = 5.5$) felt marginally more comfortable with the robot than participants who did not receive help ($M = 4.5$), $F(1, 20) = 4.0$, MSE $= 1.6$, $p < 0.06$. Participants who received help ($M = 4.5$) also felt that the responsive robot was more natural than the participants who did not receive help ($M = 2.6$), $F(1, 20) = 11.4$, MSE $= 1.8$, $p < 0.05$. Most importantly, participants who received help ($M = 4.2$) found the help provided to be more useful than participants who did not receive help ($M = 2.9$), $F(1, 20) = 5.1$, MSE $= 1.7$, $p < 0.05$. At one level it is not surprising that, when the robot helped, people found it more useful than a robot that simply nodded or did nothing at all. However, the helping behavior was a single instance of help and could have been perceived as irritating if the person had already remembered the last thing they had said. This finding was not due to the social aspects of the responsive robot; participants in the responsive robot condition ($M = 4$) did not see the robot as more useful than participants in the control condition ($M = 2.9$), $F(1, 20) = 3.4$, MSE $= 6.4$, *n.s.* Finally, note that in all cases, participants who received help had ratings above the midpoint.

### C. Discussion

This experiment had several goals. First, we wanted to explore the storytelling paradigm with a completely different population group. Along several measures, the two participant groups were quite different. In experiment 1, all of the participants were women, while in experiment 2 there was an even split of men and women. Also, the age of the participants in experiment 2 was more than twice that of experiment 1 (42 versus 20).

The results of this experiment were quite strong with respect to the success of the predictive model. We found that when the robot was running a predictive model of the person as they resumed the story, it helped people that actually needed help. For those participants who needed help in the control condition, the model accurately predicted that they would have needed help. Finally, when the model did in fact help, people found that help useful. These findings strongly suggest that the predictive model was successful.

One of the primary goals of this study was to show that a cognitive science theory based on people's memory system could be used to predict when an individual needs help resuming after an interruption. Experiment 2 used the exact same model used to match data from experiment 1 to predict when a participant would need help. The model showed success from both a technological perspective (a model of human cognition running on an embodied robotic agent) and a scientific perspective (a model predicting human performance a few seconds into the future).

The model tailored itself to individual performance and that made it somewhat sensitive to individual differences. The fact that people who got help from the robot found the robot more useful than participants who did not receive help suggests that when the robot did help the participant resume their storytelling, it was being helpful, and therefore, the predictive model was correct most of the time. The fact that some people did not need help (e.g., they resumed without any problems) suggests that if the robot had helped everyone, it would have been more irritating than helpful because people in general do not like to be offered help when it is not needed [35].

Given the single-instance nature of the task, this particular model had no real opportunities to learn. However, ACT-R's existing learning mechanisms could be applied to learn the specific parameters. The existing temporal-discounted utility learning mechanism would enable the model to learn when best to apply the productions underlying MFG. The specific model parameters used in this experiment would need to be learned over time as it gained experience.

It should also be reemphasized that this model did not take a person's social cues into account. It is very clear after listening to a person struggle to remember something that they give cues that they need help—they may have a confused look on their face, they may make a series of false starts or disfluencies or even ask explicitly for help. The overall system could likely be improved by taking these social and linguistic cues into account. It is our strong prediction that a system that is able to integrate memory, linguistic cues, and social cues would be an overall stronger model.

One other significant aspect of this experiment is that the robot had partial autonomy to perform some tasks. Its primary function—to help people resume where they left off—was an autonomous decision. It was not completely autonomous because the experimenter had to tell the robot whether the person had resumed the story. This finding is important because there are relatively few experiments with physical robots interacting with people in an autonomous manner.

## VI. GENERAL DISCUSSION

The approach that we used in this report was novel and successful. Our approach was to run a study to understand how people resume a story after an interruption (experiment 1). From that data, we built a process model of the event of interest (the resumption itself) using a computational cognitive architecture (ACT-R/E) and an existing theoretical framework (MFG). The process model was then matched to the experimental data, showing a tight fit with experimental data. This process model/experimental data fit showed that the process model is a reasonable theoretical account of how people resume a storytelling event after being interrupted. After the strong model fit, the model was placed on an embodied robotic platform and the model was applied to a completely different group of

participants and run as a predictive model (experiment 2). The predictive model did, in fact, successfully predict when people needed help after resuming a story and then provided a reminder of the last gist event that occurred before the interruption. People who received help from the embodied cognitive model were comfortable with the robotic system and found the robotic system to be more natural and more useful than participants who did not receive help.

The results from these two experiments have implications for a number of areas. First, the MFG theory we used to build the process model has traditionally been used to explain behavior and cognition on computer-based tasks [4], [8], [36]–[41]. This series of studies now suggests that the MFG model can robustly predict interruption and resumption performance in a different domain, which expands the theory's coverage.

Second, the MFG process model provides some details about how initial transactive memory occurs. Specifically, the model suggests that if people are alone, and cannot remember what they were doing, they will simply query memory or take other approaches until they eventually succeed or decide to give up. In contrast, if there is someone else available and an individual cannot remember where they were, they will ask the other person. While we did not model it in this project, it follows that the second person will occasionally rehearse the resumption point during the interruption and be able to help the interrupted individual. The rehearsal will allow them to remember the resumption point better than the interrupted person. As an individual becomes familiar with another's interests, this process may become more routine and specific so that people learn to rely on each other's knowledge and abilities.

Finally, while there are many examples of strong theories in cognitive science, relatively few of them are able to predict a phenomenon at both qualitative and quantitative levels. The current MFG model is predictive in the strong sense of the word: it not only predicts when someone has forgotten a key fact, but also changes behavior by reminding them.

This series of studies shows that running predictive models of human cognition and behavior on an autonomous platform to facilitate human–robot interaction is now a reality. ■

## REFERENCES


[1] C. Wallis, "The multitasking generation," *Time Mag.*, vol. 167, pp. 48–55, 2006.

[2] C. Thompson, "Meet the life hackers," *New York Times Mag.*, pp. 40–45, 2005.

[3] J. B. Spira, "The high cost of interruptions," *KM World*, 2005, vol. 14, no. 8.

[4] J. G. Trafton, E. M. Altmann, and R. M. Ratwani, "A memory for goals model of sequential action," *Cogn. Syst. Res.*, vol. 12, pp. 134–143, 2011.

[5] D. M. Wegner, R. Erber, and P. Raymond, "Transactive memory in close relationships," *J. Personality Social Psychol.*, vol. 61, pp. 923–929, 1991.

[6] H. M. Hodgetts and D. M. Jones, "Interruption of the tower of London task: Support for a goal activation approach," *J. Exp. Psychol., General*, vol. 135, pp. 103–115, 2006.

[7] E. M. Altmann and J. G. Trafton, "Memory for goals: An activation-based model," *Cogn. Sci.*, vol. 26, pp. 39–83, 2002.

[8] J. G. Trafton, E. M. Altmann, D. P. Brock, and F. E. Mintz, "Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal," *Int. J. Human Comput. Studies*, vol. 58, pp. 583–603, 2003.

[9] S. Y. W. Li, A. Blandford, P. Cairns, and R. M. Young, "The effect of interruptions on postcompletion and other procedural errors: An account based on the activation-based goal memory model," *J. Exp. Psychol., Appl.*, vol. 14, pp. 314–328, 2008.

[10] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin, "An integrated theory of mind," *Psychol. Rev.*, vol. 111, pp. 1036–1060, 2004.

[11] J. R. Anderson, *How Can the Human Mind Occur in the Physical Universe?* New York: Oxford Univ. Press, 2007.

[12] J. G. Trafton, M. D. Bugajska, B. R. Fransen, and R. M. Ratwani, "Integrating vision and audition within a cognitive architecture to track conversations," in *Proc. ACM/IEEE Int. Conf. Human Robot Interaction*, 2008, DOI: 10.1145/1349822.1349849.

[13] J. G. Trafton and A. M. Harrison, "Embodied spatial cognition," *Topics Cogn. Sci.*, vol. 3, pp. 686–706, 2011.

[14] A. M. Harrison and C. D. Schunn, "ACT-R/S: Look Ma, No 'Cognitive-map'!" in *Proc. Int. Conf. Cogn. Model.*, 2003, pp. 129–134.

[15] J. G. Trafton and A. M. Harrison, "Embodied spatial cognition," *Topics Cogn. Sci.*, under review.

[16] J. G. Trafton, A. M. Harrison, B. Fransen, and M. D. Bugajska, "An embodied model of infant gaze-following," in *Proc. 9th Int. Conf. Cogn. Model.*, 2009, pp. 146–151.

[17] D. D. Salvucci and N. A. Taatgen, *The Multitasking Mind*. Oxford, U.K.: Oxford Univ. Press, 2010.

[18] E. M. Altmann and W. D. Gray, "An integrated model of cognitive control in task switching," *Psychol. Rev.*, vol. 115, pp. 602–639, 2008.

[19] W. Kintsch, *Comprehension: A Paradigm for Cognition*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[20] I. McGregor and J. G. Holmes, "How storytelling shapes memory and impressions of relationship events over time," *J. Personality Social Psychol.*, vol. 76, pp. 403–419, 1999.

[21] R. C. Schank, R. P. Abelson, and R. S. Wyer, *Knowledge and Memory: The Real Story*. London, U.K.: Erlbaum, 1995.

[22] N. A. Taatgen, H. Van Rijn, and J. Anderson, "An integrated theory of prospective time interval estimation: The role of cognition, attention, and learning," *Psychol. Rev.*, vol. 114, pp. 577–598, 2007.

[23] C. Breazeal, M. Siegel, M. Berlin, J. Gray, R. Grupen, P. Deegan, J. Weber, K. Narendran, and J. McBean, "Mobile, dexterous, social robots for mobile manipulation and human-robot interaction," *Proc. ACM SIGGRAPH*, 2008, DOI: 10.1145/1401615.1401642.

[24] A. M. Harrison and J. G. Trafton, "Gaze-following and awareness of visual perspective in chimpanzees," in *Proc. 9th Int. Conf. Cogn. Model.*, 2009, pp. 270–275.

[25] W. G. Kennedy, M. D. Bugajska, A. M. Harrison, and J. G. Trafton, "'Like-Me' simulation as an effective and cognitively plausible basis for social robotics," *Int. J. Social Robot.*, vol. 1, pp. 181–194, 2009.

[26] J. G. Trafton, A. C. Schultz, D. Perznowski, M. D. Bugajska, W. Adams, N. L. Cassimatis, and D. P. Brock, "Children and robots learning to play hide and seek," in *Proc. ACM Conf. Human-Robot Interaction*, 2006, DOI: 10.1145/1121241.1121283.

[27] H. Kato, M. Billinghurst, I. Poupyrev, K. Imamoto, and K. Tachibana, "Virtual object manipulation on a table-top AR environment," in *Proc. IEEE/ACM Int. Symp. Augmented Reality*, Los Alamitos, CA, 2000, pp. 111–119.

[28] C. Nass, J. S. Steuer, and E. Tauber, "Computers are social actors," in *Proc. Conf. Human Factors Comput. Syst.*, New York, 1994, pp. 72–77.

[29] B. Mutlu, T. Shiwa, T. Kanda, and H. Ishiguro, "Footing in human-robot conversations: How robots might shape participant roles using gaze cues," in *Proc. Conf. Human Robot Interaction*, 2009, vol. 2, pp. 61–68.

[30] C. Nass and K. M. Lee, "Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction," *J. Exp. Psychol., Appl.*, vol. 7, pp. 171–181, 2001.

[31] C. F. DiSalvo, F. Gemperle, J. Forlizzi, and S. Kiesler, "All robots are not created equal: The design and perception of humanoid robot heads," in *Proc. 4th Int. Conf. Designing*

*Interactive Syst., Processes Practices Method, Tech.*, 2002, DOI: 10.1145/778712.778756.

[32] J. Goetz, S. Kiesler, and A. Powers, "Matching robot appearance and behavior to tasks to improve human-robot cooperation," in *Proc. 12th IEEE Int. Workshop Robot Human Interactive Commun.*, 2003, pp. 55–60.

[33] S. Lee, I. Y. Lau, S. Kiesler, and C. Y. Chiu, "Human mental models of humanoid robots," *Proc. IEEE Int. Conf. Robot. Autom.*, 2005, pp. 2767–2772.

[34] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, pp. 861–874, 2006.

[35] J. T. Deelstra, M. C. W. Peeters, W. B. Schaufeli, W. Stroebe, F. R. H. Zijlstra, and L. P. Van Doornen, "Receiving instrumental support at work: When help is not welcome," *J. Appl. Psychol.*, vol. 88, pp. 324–331, 2003.

[36] E. M. Altmann and J. G. Trafton, "Timecourse of recovery from task interruption: Data and a model," *Psychon. Bull. Rev.*, vol. 14, pp. 1079–1084, 2007.

[37] R. M. Ratwani, A. E. Andrews, J. D. Sousk, and J. G. Trafton, "The effect of interruption modality on primary task resumption," in *Proc. 52nd Annu. Meeting Human Factors Ergonom. Soc.*, 2008, vol. 52, no. 4, pp. 1–5.

[38] R. M. Ratwani, J. M. McCurry, and J. G. Trafton, "Predicting postcompletion errors using eye movements," in *Proc. Conf. Human Factors Comput. Syst.*, 2008, pp. 539–542.

[39] R. M. Ratwani and J. G. Trafton, "Developing a predictive model of postcompletion errors," in *Proc. 31st Annu. Conf. Cogn. Sci. Soc.*, Austin, TX, 2009, pp. 2274–2279.

[40] R. M. Ratwani and J. G. Trafton, "A generalized model for predicting postcompletion errors," *Top. Cogn. Sci.*, vol. 2, pp. 154–167, 2010.

[41] R. M. Ratwani and J. G. Trafton, "Predicting and preventing post completion errors," *Human Computer Interaction*, vol. 26, pp. 205–245, 2011.

## ABOUT THE AUTHORS

**J. Gregory Trafton**, photograph and biography not available at the time of publication.

**Allison Jacobs**, photograph and biography not available at the time of publication.

**Anthony M. Harrison**, photograph and biography not available at the time of publication.