Biologically Inspired Cognitive Architectures (2012) 1, 23–31

Available at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/bica

RESEARCH ARTICLE

# An explanatory reasoning framework for embodied agents ☆

**Laura M. Hiatt** *, **Sangeet S. Khemlani, J. Gregory Trafton**

*Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, United States*

Received 7 March 2012; received in revised form 23 March 2012; accepted 23 March 2012

**KEYWORDS**
Explanations;
Inconsistency detection;
Mental simulation;
Explanatory reasoning

**Abstract**
Our interest is in developing embodied cognitive systems. In the majority of work on cognitive modeling, the focus is on generating models that can perform specific tasks in order to understand specific reasoning processes. This approach has traditionally been exceptionally successful at accomplishing its goal. The approach encounters limitations, however, when the cognitive models are going to be used in an embodied way (e.g., on a robot). Namely, the models are too narrow to operate in the real world due to its unpredictability. In this paper, we argue that one key way for cognitive agents to better operate in real-world environments is to be able to identify and explain unexpected situations in the world; in other words, to perform explanatory reasoning. In this paper, we introduce a framework for explanatory reasoning that describes a way for cognitive agents to achieve this capability.
Published by Elsevier B.V.

## Introduction

Embodiment poses several significant challenges and opportunities for developers of cognitive agents. Typically, cognitive models are of the form of process descriptions of highly constrained experiment tasks. Chief among the challenges, therefore, is to provide agents with the ability to handle inaccuracies in its model of the ever-changing, unpredictable world. A reliable embodied cognitive agent must be able both to recognize when its model of the world is incorrect, and to explain the discrepancy in order to come up with a new model that accommodates it; i.e., it must be able to perform *explanatory reasoning*. Embodied cognition also, however, affords extraordinary opportunities for researchers in cognitive science. A prime example is the ability to interact with the world to confirm or disconfirm potential explanations of anomalies. While the above issues can potentially be relevant to a variety of non-embodied situations, we believe they are especially pertinent to embodied cognitive agents.

Consider the task of how a Navy firefighter puts out fires on board a ship. The protocol of how to put out a fire is

☆ This work was supported by Office of Naval Research Grants N0001412WX30002 and N0001411WX20516 to the first and third authors and a National Research Council Postdoctoral Research Associateship to the second author. The views and conclusions contained in this document do not represent the official policies of the US Navy.

Corresponding author. Tel.: +1 202 404 4946.
*E-mail address:* laura.hiatt@nrl.navy.mil (L.M. Hiatt).

2212-683X/$ - see front matter Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.bica.2012.03.001

fairly well defined; for instance, these are the typical steps to carry out for a two-agent team to successfully engage in extinguishing a fire in a ship compartment (Navy, 1999, chap. 555):

> First, the leader checks the door to see whether it is hot. Next, the leader opens the door and see what type of fire it is, and tells the sprayer. Then, the leader and the sprayer proceed into the compartment and fight the fire.

This procedure is a guideline, and it often suffices for most situations. But human firefighters are not bound by the guidelines, and if they detect something unexpected occurring, they are likely to adapt the procedure to accommodate it. To illustrate, imagine that a firefighting team was told that a fire was fairly moderate before entering an affected compartment, but saw tongues of fire near the ceiling and felt a rapid build-up of heat. These are two signs of an impending flashover, a very dangerous situation where all or most of the exposed combustible material in the compartment suddenly ignites. When firefighters see the two signs of the flashover, therefore, they will likely depart from the standard procedure and leave the compartment immediately.

There are at least two major challenges to creating a robotic cognitive system that can exhibit this type of flexibility. First, the robot must be able to create *expectations* of how the world should behave, and detect when the world violates those expectations (e.g., when its model of the world is inaccurate). We refer to this step as *inconsistency detection*. In the context of the above example, this means that the robot should have an idea of how it expects the fire to behave; i.e., as a relatively moderate fire typically does. Then, it should notice that the fire is not behaving this way and determine that is has found an inconsistency between its model of the world, and the world itself.

Second, the robot must be able to explain what makes this situation different, allowing it to utilize the explanation when it decides what to do next. There are several potential components to this step: (1) coming up with an idea of what may be different (e.g., generating an alternate possible model of the world); (2) elaborating on that idea (e.g., running a simulation with that model); and (3) checking whether the idea explains the current situation (e.g., seeing whether the new model accommodates the inconsistency). We refer to these three steps, which may be used in various combinations to explain inconsistencies, as *explanatory simulation*. Returning to the above example, this means that the robot should, for example, create an alternate model where the fire is not at all moderate and a flashover is imminent, realize that the model predicts that the robot will see tongues of fire and feel a rapid build-up of heat, and conclude that, since that model is consistent with what the robot knows about the world, it explains what the robot is seeing. Then, it can use its explanation to update its knowledge and adjust its course of action to leave the compartment immediately.

In this paper, we attempt to shed light on how these two extraordinarily complex challenges can be tackled to create cognitive agents that can elegantly handle the unexpected events that unfailingly occur real-world situations. Our goal in this paper is not to present an all-encompassing implementation; instead, we hope to outline and discuss the ideas key to such a system. Along those lines, we compose our discussion in the context of a framework of explanatory reasoning that performs inconsistency detection and explanatory simulation. While the tenets of this framework are rooted in psychology, they are also underspecified and can be implemented in numerous ways. Ultimately, we hope that it helps to guide interested creators of embodied cognitive systems in their quest to provide their agents with the ability to perform explanatory reasoning.

In the next section, the paper begins by briefly reviewing and discussing the extensive body of psychology research on the various components of explanatory reasoning. Then, we describe our framework of these cognitive mechanisms, and highlight two working instantiations of the framework. Finally, we end with a general discussion of the framework.

## The psychology of inconsistency detection and explanatory simulation

The psychological background of inconsistency detection and explanatory simulation is rich and complex, and has been approached from a number of different angles. Here, we discuss some of this work.

### Detecting inconsistencies

When a person's expectations conflict with their observations, it is relatively easy for them to detect that they have come across an inconsistency; this happens on a regular basis, for example, during scientific reasoning (Trickett, Trafton, & Schunn, 2009). People also, however, can be notoriously bad at detecting inconsistencies. One study by Otero and Kintsch (1992) showed that adults systematically fail to detect inconsistencies in passages such as the following (p. 230, italics added to highlight inconsistent sentences):

> Superconductivity is the disappearance of resistance to the flow of electric current. *Until now it has only been obtained by cooling certain materials to low temperatures near absolute zero*. That made its technical applications very difficult. Many laboratories are now trying to produce superconducting alloys. Many materials with this property, with immediate technical applicability, have recently been discovered. *Until now superconductivity has been achieved by considerably increasing the temperature of certain materials*.

Upon investigation, the data revealed that individuals who did not detect the inconsistency often either only recalled one of the conflicting sentences, or recalled both but discounted one of them by explaining the inconsistency away. Other studies likewise found that people err systematically in their ability to detect inconsistencies (Johnson-Laird, Legrenzi, Girotto, & Legrenzi, 2000). These studies raise a plethora of interesting questions about inconsistency detection; the major question is, of course, ''What are the mechanisms that allow people to detect inconsistencies in the world?''

In an early study on inconsistency detection, Markman (1979) gave children passages with logical inconsistencies of two types: *explicit*, such as the door was both open and not open; and *implicit*, such as the door was both open

and closed. Markman's goal was to examine the children's ability to detect inconsistencies in the text. She found that children were more likely to notice explicit rather than implicit inconsistencies. Up to the age of 12, the study showed that children failed to detect some of what appeared to be obvious implicit inconsistencies in the text, even though they were able to recall the information presented, draw deductive inferences from the text, and effectively query the experimenter. Markman took these results to suggest that to notice an inconsistency, children must carry out several demanding tasks: (1) they have to encode the information they read; (2) draw inferences from it; (3) maintain these inferences in working memory; and (4) compare their inferences with what they read. She concluded that the complexity of this process is such that children under 12 do not often carry out these procedures. To us, this study suggests that there is an explicit step to generating an expectation of the world, corresponding to her step (2) above, as well as an explicit step to using those expectations to detect inconsistencies, corresponding to her step (4).

Several studies suggest also that the complexity of generating expectations is high, and reveal important factors that help or hinder the ability to detect inconsistencies in both children and adults. These factors include the type of inconsistency (e.g., explicit falsehoods, or textual contradictions) (Baker, 1985), the relative importance of the inconsistency (Baker & Anderson, 1982;Glenberg, Wilkinson, & Epstein, 1982; Vosniadu, Pearson, & Rogers, 1988; Zabrucky & Ratner, 1986, prior beliefs (Otero & Kintsch, 1992), and instructions given before the task (Markman, 1979; Markman & Gorin, 1981). While these studies shed important light on the area of inconsistency detection, they do not offer a specific mechanism for it.

One cognitive mechanism of inconsistency detection that has been studied is based on the construction of mental models (Johnson-Laird, Girotto, & Legrenzi, 2004). A mental model is a set of entities, properties, and relations that represent a situation of what the individual believes to be true about the world. People appear to have difficulty maintaining multiple models in working memory, and so they tend to update their initial model with new information that is learned. For example, suppose an agent is told: ''all fires in compartment A are easy to put out.'' A mental model of the assertion would represent a small number of fires, all of which are tagged with the property of being easy to put out. Fig. 1 shows a schematic example of such a model. Now, if the agent were to learn something new such as ''A fire in compartment A is difficult to put out,'' it would not be able to integrate this information into its existing mental model,

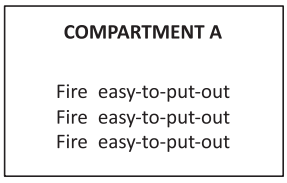


**Fig. 1** Example mental model, representing the assertion ''all fires in compartment A are easy to put out.'' Each line represents a separate individual fire that is easy to put out.

since the new statement conflicts with the previous assertion that all fires in compartment A are easy to put out. The theory posits that when a combined model cannot be constructed in this manner, people detect an inconsistency. This manner of inconsistency detection predicts the erroneous inferences about consistency in the preceding paragraphs (Johnson-Laird et al., 2000), as well as other errors in reasoning about consistency (e.g., Legrenzi, Girotto, & Johnson-Laird, 2003).

## Explanations as mental simulations

Once an inconsistency is detected, psychological evidence suggests that people next try to resolve them via explanation (Khemlani & Johnson-Laird, 2012). The psychological literature on how individuals create and judge explanations is vast (see Keil, 2006; Lombrozo, 2006, for reviews); here, we focus on research investigating explanations in response to conflicting or inconsistent information (Johnson-Laird et al., 2004; Khemlani & Johnson-Laird, 2011; Legare, Gelman, & Wellman, 2010). In these cases, there is strong evidence indicating that people explain inconsistencies by performing mental simulations to try to resolve the conflict (Trickett & Trafton, 2007).

There is a broad base of psychological and cognitive research supporting the idea of mental simulation, much of it outside the realm of explanation (Barsalou, 1999). Some of the earliest evidence on mental simulation comes from research on the mental rotation of objects and figures. Early studies began by showing that the time it takes to mentally rotate an object is proportional to the degree of rotation (Cooper & Shepard, 1973; Shepard, 1978; Shepard & Metzler, 1971), strongly suggesting that people were actually simulating the act of rotation in their mind; studies in neuroscience also suggest this (Georgopoulos, Lurito, Petrides, Schwartz, & Massey, 1989). Analogous effects occur when individuals scan a mentally encoded two-dimensional map: they take longer to estimate the distance between two places on the map as the distance between those places increases (Kosslyn, Ball, & Reiser, 1978), and they take longer to envision a journey of the map as the length of the journey increases (Bower & Morrow, 1990). When individuals comprehend text and discourse, they likewise exhibit traces of mental simulation processes ( Kelter, Kaup, & Claus, 2004; van der Meer, Beyer, Heinze, & Badel, 2002; Zwaan, 1996). For example, people are slower to understand text when they need to simulate a causal link between two sentences because one is not provided (Singer, Halldorson, Lear, & Andrusiak, 1992), and they take longer to comprehend continual events than discrete ones (Coll-Florit & Gennari, 2011).

One key study of using mental simulation for higher-level reasoning was done by Trickett & Trafton (2007). They showed strong evidence for a three-step process of hypothetical, ''what-if'' reasoning in an in-vivo study of expert scientists:

1. Generate a new representation of a system or mechanism.
2. Transform that representation in a hypothetical manner.
3. Look at the result of the simulation.

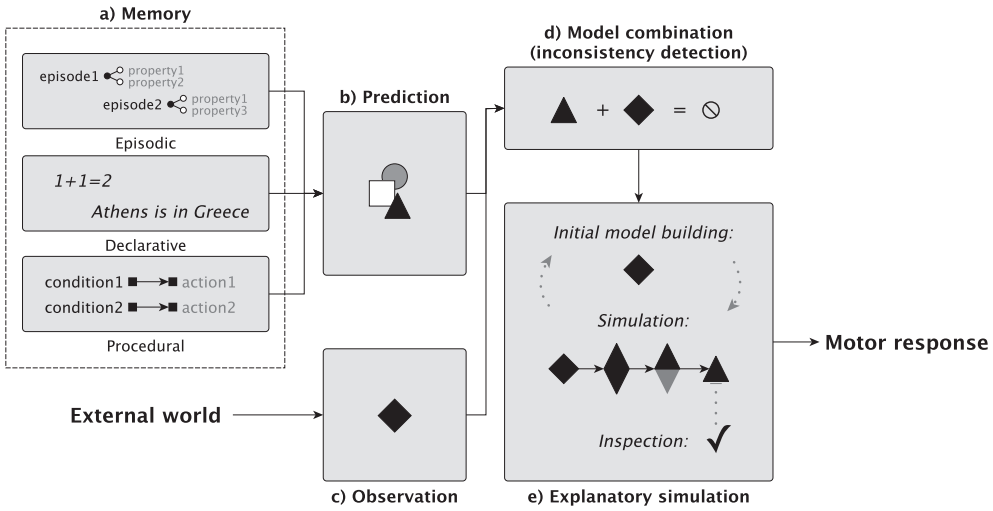


**Fig. 2** A schematic diagram of the framework for explanatory reasoning. The agent's episodic, declarative and procedural memory systems (box (a)) take data from the external world and make sense of it. Based on this, they generate expectations (box (b)) and represent observations of the world (box (c)) in a meaningful way. The expectations and observations can then be compared. If, when combined, they yield a consistent model of the world, explanatory reasoning does not need to proceed further. If, however, the system cannot reconcile its predictions with its observations, indicating that its model of the world is incorrect, it detects an inconsistency (box (d)), and tasks explanatory simulation with trying to explain the inconsistency (box (e)). Once an explanation has been found, the agent can act upon it as appropriate given its current task. It is worth it to add that memory is affected by every other component in the diagram; for clarity, we omit the arrows designating that.

For example, an astronomer might say, while reasoning with another astronomer:

> In a perfect sort of spider diagram [Step 1].
> If you looked at the velocity contours without any sort of streaming motions, no, what I'm trying to say is, um, in the absence of streaming motions [Step 2].
> You'd probably expect these lines here [gestures] to go all the way across, you know, the ring [Step 3].

The experimenters found that the scientists were most likely to engage in this type of reasoning when they were outside their own area of expertise, or on the boundary of their current knowledge. Overall, this work shows one way in which people use simulations to construct explanations in complex domains.

## An embodied framework of explanatory reasoning

In this section, we describe our framework for explanatory reasoning. A schematic diagram of the framework is given in Fig. 2. It depicts the five main components present in an agent with explanatory reasoning: (a) memory, or knowledge, of its task, the world, etc.; (b) expectations derived from its knowledge; (c) observations of the world, translated into a form that the agent understands; (d) inconsistency detection given expectations and observations; and (e) explanatory simulation. Our current focus is on parts (b), (d) and (e); (a) and (c) we in large part leave to others, for now.

### Expectations and inconsistency detection

In our framework, expectations are made explicitly, as is suggested by Markman (1979). There are many ways for an agent to generate expectations based on its knowledge. Here, in order to clearly discuss some of the ways to derive expectations, we categorize knowledge into three types; this classification, however, is not essential to our approach. *Episodic memory* describes the type of memory used to remember personally experienced events and their associated contexts, such as the time and place in which they took place; for example, that the last time an agent fought a fire in compartment A it was easily put out. *Declarative memory* describes the memory system used to remember facts and general knowledge, such as the fact that a class A fire means that only ordinary combustible material is burning.[1] *Procedural memory*, in turn, describes a system for remembering how to do things, such as how to aim and discharge a fire extinguisher. An agent's model of the world is a function of these memory systems, the specifics of which depends on, among other things, the task at hand and one's theoretic perspective. With that caveat, for clarity in this paper we will treat the agent's world model simply as the interaction of these three memory systems.

Expectations are then built from the agent's understanding of the world. An agent can, for example, expect that a current episode will unfold similarly to a previous one; for example, the firefighting robot might expect that, since the last fire in compartment A was easy to extinguish, the next one will be, as well. Declarative memory can, and should, also factor in. For example, if the robot knows that a large tank of gas was recently moved into compartment A, their expectation of the ease of fighting that fire should be adjusted accordingly. Procedural knowledge, in turn, also creates expectations rather fluidly: an agent can expect

[1] Others researchers sometimes use the term ''semantic'' memory here; we use ''declarative'' in order to be consistent with the cognitive architecture we employ, ACT-R (see the section 'Instantiations of explanatory reasoning').

that the outcome of its actions is typically the same (e.g., each time they pull the trigger of their fire extinguisher, it discharges); more interestingly, it can apply its own procedural knowledge to other agents in order to create expectations of their behavior (e.g., if the agent itself leaves the room when it suspects a flashover, it may expect that other agents will, too).

Note that this approach implies that the complexity of inconsistency detection is high. Computationally speaking, generating expectations becomes intractable fairly quickly if guidelines are not used to narrow the context of interest. This is in line with the cognitive literature; see the section 'Detecting inconsistencies' for the discussion. To illustrate, consider a scientist with a lot of knowledge about the world (or, at least, about their area of expertise). Such an expert could generate virtually a limitless number of expectations of the world; and yet, they are typically able to generate instead those which are most pertinent to the task at hand. It is left up to the individual developer to address this issue in a manner appropriate to their model.

Once an agent has an expectation, it can compare it to its observations; again, this is in line with the results suggested by Markman (1979). If an agent observes something in the world that is consistent with its expectations, it can safely integrate that observation into its memories to update its model of the world, and can carry on with its duties. Otherwise, it infers that its model of the world is inaccurate, and begins to try to explain the inconsistency.

## Explanatory simulation

Explanations are constructed by running simulations in a cognitively-plausible, three-step process (see box (e) in Fig. 2), akin to the simulation processes that humans go through during hypothetical, ''what-if'' reasoning (Trickett & Trafton, 2007). First, an initial, alternative model of the world is identified; one would envision that model to be highly similar to the agent's current model of the world, but with differences (or a key difference) that may shed light on the current inconsistency. For example, in our fire-fighting example, the agent's initial alternative model may be a world in which a flashover is about to occur. It is important to note, here, that the model is not a complete model of the world; the implications of the key difference(s) have not yet been explored (i.e., it is not yet known at this point whether the alternative model explains the agents observations). The initial model then acts as the starting state of a single simulation. Simulation is done by systematically manipulating the model to explore the implications of the differences between it and the agents current model of the world. As the simulation unfolds, it will be repeatedly inspected to see if it aligns with the agent's observations of the world. Once it is clear whether or not this is the case, the simulation stops. Obviously, if the model was found to explain the detected inconsistency (such as the above model, which explains the agents observations of fingers of fire near the ceiling), it can be folded into the current agent's model and the agent can utilize the expectation as it sees fit. Otherwise, this process can be repeated, ideally with the failure of the previous simulation helping to shed light on what a new, different, starting state should be.

It is worth noting that, if the agent has some other, easier way of generating the explanation, they should feel free to use it. For example, if this exact situation has come up before, the agent can potentially re-use the same explanation they used previously. Simulation, much like in Trickett & Trafton (2007), is envisioned here to explain inconsistencies that cannot be resolved in such a way.

To be clear, this section is not intended to describe ways to instantiate this framework. Instead, we have outlined key

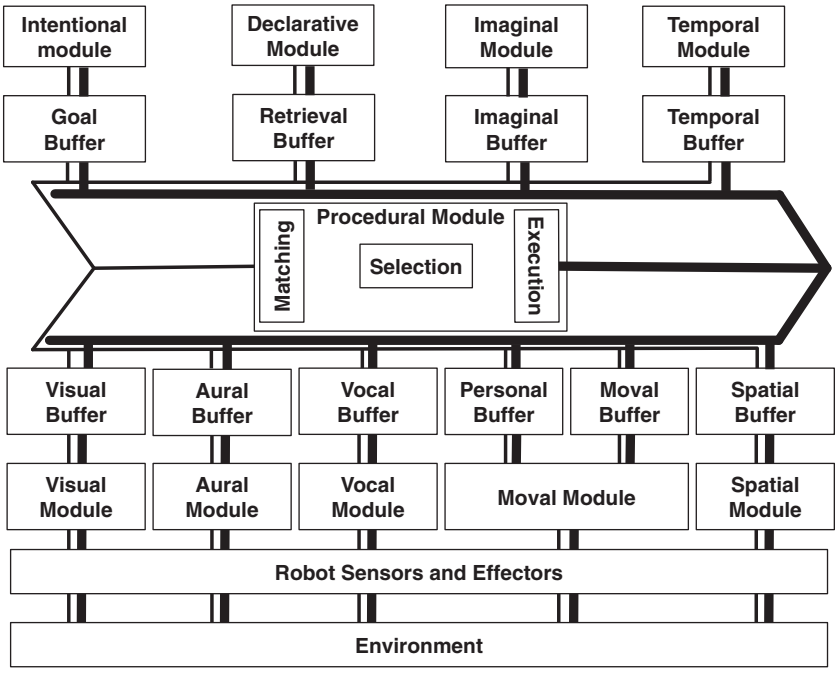


**Fig. 3** The embodied ACT-R architecture.

ideas that can be used to guide developing embodied systems capable of explanatory reasoning irrespective of differences in representation, implementation, domain, platforms, etc. We demonstrate the framework's generality next, by describing two different ways of instantiating it.

## Instantiations of explanatory reasoning

In this section, we describe two examples of explanatory reasoning in the field of cognitive robotics, and show how they represent different ways of instantiating our framework. The embodied cognitive systems in our lab are built on robotic platforms that come with various sensors for perceiving the external world and various actuators for affecting the world. We avoid a detailed discussion of our overall robotic system, except to say that we use the cognitive architecture known as ACT-R (Adaptive Character of Thought-Rational) (Anderson et al., 2004) as the basis for our robotic architecture. ACT-R is a hybrid symbolic/subsymbolic production-based system that consists of a number of modules, buffers and a central pattern matcher. Specifically, we use ACT-R's embodied configuration, ACT-R/E (Trafton & Harrison, 2011), which is used on robotic (and simulated robotic) agents (Fig. 3). ACT-R/E interfaces with world via the visual, aural, motor and vocal modules. Other modules include the intentional, imaginal, temporal and declarative modules.

The first example involves a robot who can perform theory of mind (Hiatt, Harrison, & Trafton, 2011). The second demonstrates a robot that can detect and explain episodic inconsistencies in its interactions with people. These examples utilize the Mobile, Dexterous, Social (MDS) robot (Breazeal et al., 2008) (Fig. 4), an expressive, humanoid robot which was designed for human–robot interaction.

**Fig. 4** The MDS robot.

### Theory of mind

Theory of mind refers to the ability understand the beliefs, desires and intensions of others. It is a hotly studied cognitive ability, and in recent years, developmental psychologists (Wellman, Cross, & Watson, 2001), cognitive neuroscientists (Gallese & Goldman, 1998), and cognitive modelers (Friedlander & Franklin, 2008; Hiatt & Trafton,2010) have proposed and modeled mechanisms underlying theory of mind. Hiatt et al. (2011) describes an implemented cognitive robotic system capable of carrying out theory of mind tasks, and the system can be cast as a specific instantiation of the framework we describe above. An example scenario is:

1. A robot is getting ready to go fight a fire.
2. Sunny walks up, and tells the robot that the fire in compartment A is out but that there is a new, secondary fire in compartment B.
3. The robot acknowledges this, and Sunny heads to compartment B.
4. Laura arrives and suggests that they head to compartment A to fight the fire there.
5. The robot infers that Laura does not know that the fire in compartment A is out, and tells her about it and the new fire in compartment B.
6. Laura suggests that they head to compartment B, and they both leave.

The robot's behavior is achieved by first making an expectation based on its knowledge of the world that Laura will want to head to compartment B to fight the fire there. Expectations are based on the default assumption that, given the same model, humans will behave as the robot behaves; i.e., that Laura and the robot are ''executing'' the same cognitive model. Thus, when Laura suggests heading to compartment A, the robot detects an inconsistency since it itself would head to compartment B at that time. Explanations are generated by running simulations of separate speculative cognitive models of the human, each of which differs in its knowledge about the world. Ultimately, the robot identifies which model is most likely to lead to the human's observed behavior, and uses that as the explanation of the inconsistency; here, that Laura must not know about Sunny's recent update.

One of the noteworthy things about this model is the degree to which it maintains cognitive plausibility. Simulation during theory of mind, for example, has been supported by various studies (Gallese & Goldman, 1998; Hiatt & Trafton, 2010). Furthermore, Hiatt et al. (2011) describe a human subjects experiment that also supports cognitive plausibility. In the experiment, participants were asked to rate a robot with theory of mind, a robot that points out inconsistencies but does not explain them, and a robot that neither notices inconsistencies nor explains them. Participants found the theory of mind robot to be both significantly more natural and significantly more intelligent than the other two robots.

### Detecting and explaining episodic inconsistencies

A second instantiation of our framework concerns a robot whose job is to control entry to a secure area of a building. The area is secured according to the following rule: only

employees (defined as people who carry a blue identification badge) are allowed to pass. Other persons, such as a visitor (defined as a person who carries an orange identification badge) are not allowed to pass unless they are escorted by an employee. In this cognitive system, the robot is able to remember typical episodes of interactions with people, and detect and explain any inconsistencies between a relevant typical episode and the current episode. Although ACT-R does not have an explicit theory of episodic memory, here we consider it to be part of its declarative memory system. Consider the interaction below; here, Laura is an employee who typically wears a blue badge:

1. Laura walks up to the robot; she is not wearing an employee's badge, but rather an orange visitor's badge.
2. The robot greets Laura, and processes the color of her badge.
3. The robot hypothesizes that Laura forgot her normal badge, and so she has to wear a temporary visitor's badge.
4. Laura confirms the explanation.
5. The robot denies Laura entry, since she is not escorted by an employee with a blue badge.

Expectations are generated by retrieving relevant typical episodes from declarative memory. Here, when the robot sees Laura, it automatically accesses a typical episode of interaction with Laura from memory, and uses it as its expectation. Of course, since in this typical interaction Laura is wearing a blue badge, the robot knows it is in an anomalous situation. Explanations are then achieved by performing simulation via backwards counterfactual chaining. The simulation creates two chains of procedure sequences in parallel: one with a starting state of the expected episode of an interaction with Laura; the other, with a starting state of the current episode. The robot searches backwards in a chain by applying procedures in ''reverse''; for example, given a starting state of Laura walking up to the robot wearing her typical employee badge, the robot may apply a ''reversed'' procedure of Laura leaving for work to create a new state of Laura getting into her car with her employee badge. The robot does this for both chains to find the point at which they intersect, and uses it as the basis of its explanation. Here, the point of intersection (Laura getting ready to go to work) is immediately followed by a procedure where Laura remembers her badge (for the expected episode), versus a procedure where Laura forgets her badge (for the current episode), allowing the robot to infer that today, Laura forgot her badge.

This system is fully implemented and runs on our MDS robotic platform. The system assumes that the robot has enough knowledge of the world to perform explanatory reasoning; this knowledge, however, is dynamic and as time goes on the robot's expectations of people, and explanations of inconsistencies, will likely change.

## General discussion

The systems we develop are cognitive models that operate on robotic platforms, and so they need to be able to sense, communicate with, and act on their environment. As we stated in the very first paragraph of this paper, embodiment is both a challenge and an opportunity. The reasons for its difficulty are perhaps the more intuitive ones: the world, and the agents in it, are unpredictable, and a cognitive agent operating in it needs to be able to make sense of conflicting, uncertain, and incorrect information. In the context of this paper, this means that embodied agents need the ability to detect and explain inconsistencies in the world.

So how can embodied agents cope with situations that violate their expectations? We believe that the capability of explanatory reasoning will allow them to do so. In this paper, we have introduced a framework for explanatory reasoning. The framework is a way of characterizing the classes of cognitive and computational mechanisms that can be used to perform inconsistency detection and explanatory simulation. This framework is underspecified; it can be instantiated to conduct many specific tasks, such as in the two robotic systems we describe above. It posits that agents construct expectations that are compared to observations in the external world. When an inconsistency is detected, the framework explains the conflict via explanatory simulation. Once an adequate explanation is constructed, the agent is free to act on the explanation to seek more information, change underlying beliefs, or adapt its behavior, as it sees fit.

The second point, which we have largely glossed over thus far, is that embodied, cognitively plausible systems offer unique opportunities to researchers both in cognitive science and robotics. To loosely borrow a term from biology, we like to think of cognitive robotics as a symbiotic relationship between cognitive scientists and roboticists, where both parties could certainly survive apart, but both can enjoy great benefits from living together. For cognitive scientists, embodied cognitive models provide a new set of tools to capitalize on to more fully understand human cognition. Specifically, experiments involving cognitive models are no longer limited to the stationary computer, and can, in addition, utilize cognitive models that can act upon the world.

Looking at this from a robotics point of view, cognitive science provides robots with tools that allow them to better interact with humans. The use of similar representations, for instance, maximizes communicative fidelity because it ensures that one agent can transfer a representation to another agent without losing the relevant relationships that make the representation informative (Kurup, Bignoli, Scally, & Cassimatis, 2011; Kurup, Lebiere, Stentz, & Herbert, 2012). As another example, cognitive robots have been shown to be better partners because they can capitalize on the knowledge that cognitive models of their human counterparts can provide them (Hiatt et al., 2011; Kennedy, Bugajska, Harrison, & Trafton, 2009; Trafton et al., 2006).

Along those lines, although we have approached our framework keeping cognitive plausibility at the forefront, it is not strictly necessary to adopt our approach in a cognitively plausible way. Certainly, each aspect of our approach is deeply rooted in psychological literature, such as the use of mental simulations to resolve inconsistencies. The specificities of a working instantiation of the framework, however, are left to the developer and can be as psycholog-

ically plausible or implausible as they desire. For example, a researcher interested in cognitive plausibility may implement mental simulation in a cognitively plausible way, such as by limiting it to piecemeal transformations as suggested by Hegarty (1992, 2004); one less interested in plausibility may decide to take advantage of a robot's computational power and use a full robotic simulation environment to perform its mental simulations. Both routes are productive in their own way.

We end with a note about the framework's generality. Many of the key components of the framework are general mechanisms, whose ideas can be applied to any number of domains, architectures, robotic platforms, representations, etc. We do expect that different implementations of simulation, for example, will place additional constraints on it (for example, an implementation of simulation for an ACT-R cognitive model will look very different from an implementation of simulation for the mental model theory), but, at a high level, they will look the same. The same can be said of comparing an expectation with an observation to determine whether there is an inconsistency. In contrast, we expect that how explanations are used will be very domain-dependent: for example, in the fire-fighting domain, one may want to rush out of a compartment if a flashover is considered a possible explanation; but in a different domain it may be appropriate for an agent to take steps to confirm an explanation before reacting to it so drastically. Such interactions between specific and general mechanisms are extremely interesting, and we look forward to exploring this area more.

## Acknowledgement

Special thanks to Anthony Harrison for his helpful comments and discussion on this research.

## References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*, 1036–1060.

Baker, L. (1985). How do we know when we don't understand? Standards for evaluating text comprehension. In D. Forrest-Pressley, G. Mackinnon, & T. Waller (Eds.), *Metacognition, cognition, and human performance*. New York, NY: Academic Press.

Baker, L., & Anderson, R. I. (1982). Effects of inconsistent information on text processing: Evidence from comprehension monitoring. *Reading Research Quarterly, 17*, 281–294.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577–609.

Bower, G. H., & Morrow, D. G. (1990). Mental models in narrative comprehension. *Science, 247*, 44–48.

Breazeal, C., Siegel, M., Berlin, M., Gray, J., Grupen, R., Deegan, P., et al. (2008). Mobile, dexterous, social robots for mobile manipulation and human–robot interaction. In *SIGGRAPH '08: ACM SIGGRAPH 2008 new tech demos*.

Coll-Florit, M., & Gennari, S. P. (2011). Time in language: Event duration in language comprehension. *Cognitive Psychology, 62*, 41–79.

Cooper, L. A., & Shepard, R. (1973). Chronometric studies of the rotation of mental images. In W. G. Chase (Ed.), *Visual information processing*. New York, NY: Academic Press.

Friedlander, D., & Franklin, S. (2008). LIDA and a theory of mind. In P. Wang, B. Goertzel, & S. Franklin (Eds.), *Artificial general intelligence* (pp. 137–148). IOS Press.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2.

Georgopoulos, A. P., Lurito, J. T., Petrides, M., Schwartz, A. B., & Massey, J. T. (1989). Mental rotation of the neuronal population vector. *Science*, 243.

Glenberg, A., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition, 10*, 597–602.

Hegarty, M. (1992). Mental animation: Inferring motion from static displays of mechanical systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 1084–1102.

Hegarty, M. (2004). Mechanical reasoning as mental simulation. *Trends in Cognitive Sciences, 8*, 280–285.

Hiatt, L. M., & Trafton, J. G. (2010). A cognitive model of theory of mind. In *Proceedings of the international conference on cognitive modeling*.

Hiatt, L. M., Harrison, A. M., & Trafton, J. G. (2011). Accommodating human variability in human–robot teams through theory of mind. In *Proceedings of the international joint conference on artificial intelligence 2011*.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., & Legrenzi, M. (2000). Illusions in reasoning about consistency. *Science, 288*, 531–532.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review, 111*, 640–661.

Keil, F. (2006). Explanation and understanding. *Annual Review of Psychology, 57*, 225–254.

Kelter, S., Kaup, B., & Claus, B. (2004). Representing a described sequence of events: A dynamic view of narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 451–464.

Kennedy, W. G., Bugajska, M. D., Harrison, A. M., & Trafton, J. G. (2009). ''Like-me'' simulation as an effective and cognitively plausible basis for social robotics. *International Journal of Social Robotics, 1*, 181–194.

Khemlani, S., & Johnson-Laird, P. N. (2011). The need to explain. *Quarterly Journal of Experimental Psychology, 64*, 2276–2288.

Khemlani, S., & Johnson-Laird, P. N. (2012). Hidden conflicts: Explanations make inconsistencies harder to detect. *Acta Psychologica, 139*, 486–491.

Kosslyn, S. M., Ball, T. M., & Reiser, B. J. (1978). Visual images preserve metric spatial information: Evidence from studies of image scanning. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 47–60.

Kurup, U., Bignoli, P., Scally, J., & Cassimatis, N. L. (2011). An architectural framework for complex cognition. *Cognitive Systems Research*, 12.

Kurup, U., Lebiere, C., Stentz, A., & Herbert, M. (2012). The importance of shared mental models and shared situation awareness for transforming robots from tools to teammates. In *Proceedings of the SPIE conference on defense, security, and sensing* (Vol. 3837).

Legare, C. H., Gelman, S. A., & Wellman, H. M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development, 81*, 929–944.

Legrenzi, P., Girotto, V., & Johnson-Laird, P. N. (2003). Models of consistency. *Psychological Science, 14*, 131–137.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences, 10*, 464–470.

Markman, E. (1979). Realizing that you don't understand: Elementary school children's awareness of inconsistencies. *Child Development, 50*, 643—655.

Markman, E., & Gorin, L. (1981). Children's ability to adjust their standards for evaluating comprehension. *Journal of Educational Psychology, 73*, 320—325.

United States Navy. (1999). Surface ship firefighting. In *Naval Ship's technical manual* (Vol. 1).

Otero, J., & Kintsch, W. (1992). Failures to detect contradictions in text: What readers believe versus what they read. *Psychological Science, 3*, 229—235.

Shepard, R. (1978). The mental image. *American Psychologist, 33*, 125—137.

Shepard, R., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*, 701—703.

Singer, M., Halldorson, M., Lear, J. C., & Andrusiak, P. (1992). Validation of causal bridging inferences in discourse understanding. *Journal of Memory and Language, 31*, 507—524.

Trafton, J. G., & Harrison, A. M. (2011). Embodied spatial cognition. *Topics in Cognitive Science, 3*, 686—706.

Trafton, J. G., Shultz, A. C., Cassimatis, N. L., Hiatt, L. M., Perzanowski, D., Brock, D. P., et al (2006). Communicating and collaborating with robotic agents. In R. Sun (Ed.), *Cognition and multi-agent interaction* (pp. 252—278). Cambridge University Press.

Trickett, S., & Trafton, J. G. (2007). ''What if...'': The use of conceptual simulations in scientific reasoning. *Cognitive Science, 31*, 843—875.

Trickett, S., Trafton, J. G., & Schunn, C. (2009). How do scientists respond to anomalies? Different strategies used in basic and applied science. *Topics in Cognitive Science, 1*, 711—729.

van der Meer, E., Beyer, R., Heinze, B., & Badel, I. (2002). Temporal order relations in language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 770—779.

Vosniadu, S. P., Pearson, D. P., & Rogers, T. (1988). What causes children's failure to detect inconsistencies in text? Representation versus comparison difficulties. *Journal of Educational Psychology, 80*, 27—39.

Wellman, H. W., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*, 655—684.

Zabrucky, K., & Ratner, H. H. (1986). Children's comprehension monitoring and recall of inconsistent stories. *Child Development, 57*, 1401—1418.

Zwaan, R. (1996). Processing narrative time shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1196—1207.