Explaining Task Processing in Cognitive Assistants That Learn

Deborah L. McGuinness¹, Alyssa Glass¹,², Michael Wolverton², and Paulo Pinheiro da Silva¹,³

Stanford University¹, Stanford, CA {dlm | glass}@ksl.stanford.edu
SRI International², Menlo Park, CA mjw@ai.sri.com
University of Texas, El Paso³, El Paso, TX paulo@utep.edu

Abstract
As personal assistant software matures and assumes more autonomous control of user activities, it becomes more critical that this software can explain its task processing. It must be able to tell the user why it is doing what it is doing, and instill trust in the user that its task knowledge reflects standard practice and is being appropriately applied. We will describe the ICEE (Integrated Cognitive Explanation Environment) explanation system and its application to explaining task reasoning. Key features include (1) an architecture designed for re-use among different task execution systems; (2) a set of introspective predicates and a software wrapper that extract explanation-relevant information from a task execution system; (3) a version of the Inference Web explainer for generating formal justifications of task processing and converting them to user-friendly explanations; and (4) a unified framework for explaining results from task execution, learning, and deductive reasoning. Our work is focused on explaining belief-desire-intention (BDI) agent execution frameworks with the ability to learn. We demonstrate ICEE’s application within CALO, a state-of-the-art personal software assistant, to explain the task reasoning of one such execution system and describe our associated trust study.

Introduction
Personalized software assistants have the potential to support humans in everyday tasks by providing assistance in cognitive processing. If these agents are expected to achieve their potential and perform activities in service of humans (and possibly other agents) then these agents need to be fully accountable. Before their users can be expected to rely on cognitive agents, the agents need to provide the users with justifications for their decisions, including that those decisions are based on appropriate processes and on information that is accurate and current. The agents need to be able to use these justifications to derive explanations describing how they arrived with the ability to abstract away detail that may be irrelevant to the user’s understanding and trust evaluation process. Further, if the agents are to be used to perform tasks, they need to explain how and under what conditions they will execute a task, as well as how and why that procedure has been created or modified over time.

One significant challenge to explaining cognitive assistants is that they, by necessity, include task processing components that evaluate and execute tasks, as well as reasoning components that take input and determine conclusions. Thus, a cognitive assistant explainer will need to explain task processing responses as well as more traditional reasoning systems, providing access to both inference and provenance information, which we refer to as knowledge provenance (Pinheiro da Silva, McGuinness, McCool, 2004).

Work has been done in the theorem proving community, as well as in many specialized reasoning communities, to explain deductions. A limited amount of work has also been done in the task execution community. What has not been done is work explaining task execution in a way that is also appropriate for explaining deductive reasoning and provenance. Our work provides a uniform approach to representing and explaining results (including provenance) from both communities, with the additional emphasis on providing explanations of learned information.

Our design and implementation work is in the setting of the DARPA Personalized Assistant that Learns (PAL) program, as part of the Cognitive Assistant that Learns and Organizes (CALO) project. The CALO system includes work from 22 different organizations. This presents a complex explanation challenge where CALO must explain conclusions from multiple knowledge sources, both hand built and automatically generated, with multiple reasoning techniques including task processing, deduction, and learning. In this paper, we will present our representation, infrastructure, and solution architecture for explaining BDI-based task processing; describe how it has been implemented in our new Integrated Cognitive Explanation Environment (ICEE); and show how it has been used to explain cognitive assistants. We also discuss preliminary results from a study of CALO users that show how explanation can be a key component in building user trust in cognitive assistants that learn.

Motivating Scenario
As a motivating scenario, we provide an example from the CALO office domain. In this example, the cognitive agent...
is working in the role of an office assistant, and has been tasked with purchasing a laptop for its user. In order to accomplish the high-level goal of buying a laptop, the cognitive agent uses a simple three step process with a laptop specification as input. The first subtask, GetQuotes, requires the agent to obtain three quotes from three different sources. The second subtask, GetApproval, requires a particular form to be signed by an approval organization representative. The final subtask, SendOrderToPurchasing, requires a requisition form to be sent to purchasing. Note that in sequential tasks such as these, the termination condition of a previous subtask is typically a precondition to the next subtask.

The user may ask for an explanation of the agent’s behavior at any time. ICEE provides a dialogue interface to explaining the agent’s actions. The user can request explanations starting with any of several supported explanation request types. For example, “Why are you doing <subtask>?”, is an example of an explanation request concerning task motivation. Other task explanation request types include execution history, plans, status, task ordering, or explicit questions about time. ICEE also handles extensive questions about task provenance, including explanations about the requestor of a task and the source of the procedure being executed by the system. These questions have been guided by an initial study focused on how to build user trust in cognitive agents.

ICEE contains one or more explanation strategies for each explanation request type. Based on context such as system status and past interactions, in addition to a model of the user, ICEE chooses one strategy. For example, in response to the above question, ICEE may choose the simple strategy of revealing the task hierarchy: “I am trying to do <high-level-task>, and <subtask> is one subgoal in the process.” For each explanation request, ICEE can either reuse an existing task justification, which includes a task execution trace, or build and parse a new justification. Then ICEE presents as explanations the portions of the justification that are relevant to the explanation strategy. Additionally, ICEE suggests follow-up explanation requests for the user, enabling mixed initiative dialogue between the agent and the user. For example, follow-up questions in the above scenario might include:

“Why are you doing <high-level-task>?”
“Why haven’t you completed <subtask> yet?”
“Why is <subtask> a subgoal of <high-level-task>?”
“When will you finish <subtask>?”
“What sources did you use to do <subtask>?”

The ICEE System

Architecture Overview

The architecture of ICEE is designed to be flexible and allow explanations to be derived from justifications gathered seamlessly from a variety of task processing and knowledge systems. An explanation dispatcher gathers structured explanation requests from the user through a collaboration agent or user interface. The assistant’s user interface provides specialized mechanisms for users to request explanations.

Based on the type of the explanation request, the explanation dispatcher determines which explainer will handle the request, and forwards it to the proper explainer component. For questions related to task processing, the Task Manager (TM) explainer handles the request. The TM explainer instructs the TM wrapper to gather task processing information about the requested tasks. The TM wrapper is closely coupled with a BDI execution system, or task manager. We have provided a TM wrapper for the task execution system used in CALO, which is based on the SRI Procedural Agent Realization Kit (SPARK); however, any similar task execution system could be similarly enhanced with explanation capabilities.

The TM wrapper stores the gathered task processing information in the task state database. This database is then used by the justification generator to create a justification for the tasks currently under execution, including any additional processing that is related to the current tasks. The generated justification can then be used by the TM strategies to create alternative explanations and select the one most salient to the user’s questions. The explanation dispatcher returns the selected explanation to the collaboration agent for appropriate display to the user. Each of these architectural components is discussed below.

Task Oriented Processing

Complex cognitive agents must have a mechanism for representing and executing tasks. A belief-desire-intention (BDI) model (Rao & Georgeff, 1995) is a common framework for task reasoning components. BDI systems cover a range of execution capabilities, including hierarchical task encoding, control of procedural agent behavior, sequential and parallel execution of sub-procedures, conditional execution, branching, flexible preconditions, and meta-level reasoning to determine applicable procedures.

Task management in CALO is provided by SPARK (Morley & Myers 2004), a BDI agent framework, which maintains procedures that define agent actions. The CALO Task Manager’s knowledge base includes human-authored procedures along with procedures that were partially or completely learned based on evolving information. ICEE gathers information on both static aspects of procedures within SPARK as well as dynamic information about its past and current execution state.

Introspective Predicates

ICEE is designed to provide cognitive agent users with the ability to ask detailed questions about task execution and
to engage in a mixed initiative dialogue about past, current, and future task execution. To provide detailed explanations of task processing system behavior, justifications are annotated with meta-data. In order to generate detailed explanations, the task execution system must be able to expose this meta-information. One of our contributions is a specification of a set of introspective predicates that were designed to provide access to meta-information required for explainable task processors. These introspective predicates fall into three categories:

1. **Basic Procedure Information:** relatively stable, static information that is not dependent on when a task is executed. Provenance information about how task definitions have been created or learned is a key aspect of these introspective predicates.
2. **Execution Information:** dynamic information that is generated as a task begins being executed, and remains valid in some form throughout the execution of that task. This information also includes history related to completed tasks.
3. **Projection Information:** information about future execution, as well as alternatives for decision points that have already passed.

A task execution system that provides access to this set of introspective predicates can be linked to ICEE and allow it to fully explain all the question types and strategies described above. Details on the introspective predicates can be found in (Glass & McGuinness 2006).

**Wrapper, Action Schema and Action Database**

In order to collect explanation-relevant information from the task execution agent and store it in a format understandable by the explainer, we designed and built a wrapper for SPARK and an intermediate action schema in which to record task execution information. These elements were designed to achieve three criteria:

- **Salience.** The wrapper should obtain information about an agent’s processing that is likely to address some possible user information needs.
- **Reusability.** The wrapper should obtain information that is also useful in other cognitive agent activities that require reasoning about action—for example, state estimation and procedure learning.
- **Generality.** The schema should represent action information in as general a way as possible, covering the action reasoning of blackboard systems, production systems, and other agent architectures.

The wrapper collects a snapshot of SPARK’s current state as well as the previous decisions that led to that state. It uses SPARK’s expanded introspective predicates to extract the portions of its underlying intention structure relevant to its current intentions, building this structure by recursively querying for the supporting elements of intentions and procedures. Example queries include: What are the current intentions? What is the procedure instance that led to intention \( \lambda \)? What are the preconditions that were met before procedure \( P \) could be executed?

After collecting the snapshot, the wrapper stores it in a SPARK-independent task execution action database. This schema reflects that most task execution systems share the same common structure. While the current terminology in our schema is consistent with SPARK’s, the concepts are general and consistent with other cognitive architectures. For example, “procedures” in our schema are equivalent to “knowledge sources” in BB* and other blackboard architectures, “procedure instances” are equivalent to “Knowledge Source Activation Records (KSARs)”, etc. (Hayes-Roth 1985). The database records the relationships between key entities relevant to the agent’s current state, for example, which intentions were established by which procedure instances, which procedure a given procedure instance instantiates, and which of a procedure’s termination conditions were satisfied and which were not.

We achieve multiple design goals by creating a system-specific wrapper and a generic action schema. When new task execution systems (other than SPARK) need to be explained, the generic action schema is reused and only a new wrapper is needed. No changes are required to the justification representation or the explanation strategies. Also, the action schema can be reused by other cognitive agent components for purposes beyond explanation, such as state capture, archiving, or snapshotting.

**Generating Formal Justifications**

A cognitive agent’s actions should be supported by justifications that are used to derive and present understandable explanations to end-users. These justifications need to reflect both how the actions support various user goals, and how the particular actions chosen by the agent were guided by the state of the world. More specifically, our approach to task justification breaks down the justification of a question about a particular task \( T \) into three complementary strategies, described here using terminology from SPARK:

- **Relevance:** Demonstrate that fulfilling \( T \) will further one of the agent’s high-level goals, which the user already knows about and accepts
- **Applicability:** Demonstrate that the conditions necessary to start \( T \) were met at the time \( T \) started (possibly including the conditions that led \( T \) to be preferred over alternative tasks)
- **Termination:** Demonstrate whether one or more of the conditions necessary to terminate \( T \) has not been met.

This three-strategy approach contrasts with previous approaches to explanation, most of which dealt with explaining inference (Scott et al. 1984, Wick & Thompson 1992). Previous approaches generally have not dealt with termination issues, and they also generally have not distinguished between relevance and applicability conditions. These are critical aspects of task processing and thus are important new issues for explanation.

Justifications can be seen and represented as proofs of how information was manipulated to come to a particular
Establishing Trust in Cognitive Agents

To evaluate the use and effect of explanation in cognitive agents, we conducted a preliminary trust study among a set of CALO users. We studied 10 users from SRI International who performed a variety of tasks with individual CALO agents over approximately two weeks. The users spanned a wide range of ages, backgrounds, and prior experience with CALO. Participants were informed before the study that they would be interviewed about system trust and understanding. We then interviewed each of the users, following a structured script of questions with both free-form and rank-based answers. We had two primary aims for this study: first, to identify what factors users believe influence their trust in a complex cognitive agent; and second, to identify which types of questions, if any, users would like to ask a cognitive assistant to better understand its answers.

Preliminary results from this study show that understanding system decisions, including any unexpected decisions (e.g., inconsistencies), and thus establishing trust in a system as complex as CALO, requires a multi-pronged approach. While trust in any system generally requires a series of positive interactions over an extended period of time, the complexity and constant change of an adaptive

1 http://iw.stanford.edu/2006/06/pml-provenance.owl
2 http://iw.stanford.edu/2006/06/pml-justification.owl
3 http://iw.stanford.edu/2006/06/pml-trust.owl

Establishing Trust in Cognitive Agents

To evaluate the use and effect of explanation in cognitive agents, we conducted a preliminary trust study among a set of CALO users. We studied 10 users from SRI International who performed a variety of tasks with individual CALO agents over approximately two weeks. The users spanned a wide range of ages, backgrounds, and prior experience with CALO. Participants were informed before the study that they would be interviewed about system trust and understanding. We then interviewed each of the users, following a structured script of questions with both free-form and rank-based answers. We had two primary aims for this study: first, to identify what factors users believe influence their trust in a complex cognitive agent; and second, to identify which types of questions, if any, users would like to ask a cognitive assistant to better understand its answers.

Preliminary results from this study show that understanding system decisions, including any unexpected decisions (e.g., inconsistencies), and thus establishing trust in a system as complex as CALO, requires a multi-pronged approach. While trust in any system generally requires a series of positive interactions over an extended period of time, the complexity and constant change of an adaptive

1 http://iw.stanford.edu/2006/06/pml-provenance.owl
2 http://iw.stanford.edu/2006/06/pml-justification.owl
3 http://iw.stanford.edu/2006/06/pml-trust.owl
understanding the notion of trust, and how explanation of declarative rule processing, and provenance. of task executions in the presence of learned knowledge, work further expands the coverage to support explanations on settings that are web-based or distributed. Our current explaining hybrid reasoners added to the field by focusing rules. Previous work on Inference Web and related work notion of explaining the deductive trace of declarative reasoning systems. Most of these have focused on some systems and, to a lesser extent, explaining automated. There has been an abundance of work in explaining expert trust. We are continuing to evaluate the results of our trust behaviors, can similarly be key to helping users to build cognitive agents, particularly in the presence of learned trust. We believe that explanation of task execution can help users to establish and build deductive reasoning can also be successfully applied to explanatory the breadth required by cognitive assistants.

Driven by the needs of explaining cognitive assistants, we needed to focus on explanation infrastructures that can work with task execution systems as well as with deductive reasoners and learning components. We have described current progress on designing and implementing an extensible and flexible architecture that is capable of explaining the breadth required by cognitive assistants.

We began our work by producing an enumeration and categorization of classes of cognitive assistant explanations, which was further motivated and refined using the initial results of the CALO trust study previously described. We also analyzed the existing execution environment to identify the actions taken that would need to be explained, leading to a characterization of the SPARK execution environment. We believe that this characterization is representative not just of SPARK, but also of other agent architectures, most notably BDI-based agents and blackboard-based systems. The above characterization is integrated with an explanation architecture capable of explaining standard reasoners and hybrid theorem proving environments, an approach we are further testing as part of a new system, funded by the DARPA Integrated Learning program (IL 2006).

Our architecture and implementation demonstrates that an explanation mechanism initially designed for explaining deductive reasoning can also be successfully applied to explaining task-oriented reasoning. We developed a three strategy approach to explaining task execution including relevance, applicability, and termination. Additionally,
work devoted simply to explaining task execution has not traditionally focused on explaining a broader setting including deduction and provenance. We developed a framework for extracting and representing the state and history of a task system’s reasoning, a framework that is already proving to be useful for building trust and explanation services in other agent activities.

ICEE includes initial implementations of all of the described components, and is seeded with a limited number of strategies, as prioritized by our trust study. Current work includes expanding these components to provide advanced functionality; additional capabilities for handling the results of procedure learning (from instruction and demonstration) that extend and/or update the procedures that the task execution system uses; as well as a strong focus on explaining conflicts, explaining failures, and further explaining knowledge provenance.

Conclusions

If cognitive agents are to achieve their potential as trusted assistants, they must be able to explain their actions and recommendations. In this paper, we presented our explanation infrastructure and describe our implementation for the CALO cognitive assistant. While we focused in this paper on explaining the core task processing component, our solution provides a uniform way of encoding task execution as well as deductive reasoning and learning components as task processing justifications. Additionally, the approach is integrated with the Inference Web infrastructure for supporting knowledge provenance so that explanations derived from task processing justifications may be augmented with source information, improving end-user understanding of answers. Further evaluation has demonstrated that a better understanding of answers by users is a key factor for users to establish trust in cognitive assistants. The primary contributions of this paper are (1) the design and reference implementation of a general and reusable explanation infrastructure for cognitive assistants that integrate task processing, deductive reasoning, and learning; and (2) a supporting design describing the necessary information for explaining task processing in changing environments (including the specification of the relevance, applicability, termination strategy, and the categorization of explanation requests and follow-up strategies), with a focus on changing environments, such as those being updated by learners.

References
