

Trustable Task Processing Systems

Alyssa Glass, Deborah L. McGuinness, Paulo Pinheiro da Silva, Michael Wolverton

As personal assistant software matures and assumes more autonomous control of user activities, it becomes more critical that this software can 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. Our research focuses broadly on providing infrastructure that may be used to increase trust in intelligent agents. In this paper, we will report on a study we designed to identify factors that influence trust in intelligent adaptive agents. We will then introduce our work on explaining adaptive task processing agents as motivated by the results of the trust study. We will introduce our task execution explanation component and provide examples in the context of a particular adaptive agent named CALO. 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 extracts 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.

1 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 justifications for their decisions, including that those decisions are based on appropriate processes and on information that is accurate and current. 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.

One challenge to explaining adaptive assistants is that they, by necessity, include task processing components that evaluate and execute tasks, as well as reasoning components that determine conclusions. A comprehensive explainer needs 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 [1].

Work has been done in the theorem proving community, as well as in many specialized reasoning communities, to explain deductions. A limited amount of explanation work has 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 *both* deductive reasoning and provenance. Our work provides a uniform approach to representing and explaining provenance and results from both communities, in addition to learned information.

We present our work in the setting of the DARPA Personalized Assistant that Learns (PAL) [2] program, as part of the Cognitive Assistant that Learns and Organizes (CALO) [3] project. The CALO system includes work from 22 different organizations. This presents a complex challenge where CALO users must understand and trust 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 first

describe a study of CALO users, in which we investigate issues of trust and usability, discussing several design guidelines implied by the results of this study. Using these guidelines, we then present our representation, infrastructure, and solution architecture for explaining BDI-based task processing and learning in adaptive agents. We describe how it has been implemented in our new Integrated Cognitive Explanation Environment (ICEE), and show how it has been used to provide explanations in CALO.

2 Trust Study

We conducted a structured, qualitative trust study to identify what factors influence user trust in adaptive agents, to understand the types of questions users would like supported by such a system, and to evaluate the general usability of adaptive agents.

Procedure Our study was conducted in two basic stages: the usage stage and the interview stage. For the usage stage, we piggy-backed on a broad study aimed at testing the learning capabilities within the CALO system. The broad study is part of a long term effort involving a large set of testers and researchers, extensive participant training for the use of various CALO components, and detailed analysis of complex data logs and learning algorithms through intensive system use by dedicated users over approximately two weeks. During the usage stage, participants typically used the system for a full eight hour work day, each day, for the entire duration of the test period.

During the interview stage, we interviewed each participant after the usage period. The interviews were structured to follow a fixed script for all participants. The script contained 40 questions—eleven questions using a typical 5-step Likert scale (from “hardly ever” to “extremely often”) and 29 open response questions. The script was organized around five main topics: failure, surprise, confusion, question-answering, and trust. Each interview was audio recorded. We used these recordings to make notes and to organize the responses into common themes.