Explainable Adaptive Assistants
Deborah L. McGuinness, Tetherless World Constellation, RPI
Alyssa Glass, Stanford University
Michael Wolverton, SRI International
Paulo Pinheiro da Silva, University of Texas at El Paso
Cynthia Chang, Tetherless World Constellation, RPI

Motivation
Usable adaptive assistants need to be able to explain their recommendations if users are expected to trust them. Our work on ICEE -- the Integrated Cognitive Explanation Environment -- provides an extensible infrastructure for supporting explanations. We aim to improve trust in learning-enabled agents by providing transparency concerning:

- Provenance
- Information manipulation
- Task processing
- Learning

The ICEE explainer includes:

- Descriptions of question types and explanation strategies
- Architecture for generating interoperable, machine interpretable, sharable justifications of answers containing enough information to generate explanations
- Components capable of obtaining justification information from SPARK (a BDI agent architecture)
- History of execution states as justifications

Explanation Foundation
PML: Provenance, Justification, and Trust Interlingua
Inference Web Toolkit: Suite of tools for generating, browsing, searching, validating, and summarizing explanations
ICEE: Explanation framework for explaining cognitive agents, with focus on task processing and learning

Architecture

User Study
Interviewed 13 adaptive agent users, focusing on trust, failures, surprises, and other sources of confusion. Identified themes on trusting adaptive agents, including:

- Learning can make users feel ignored
- Users want to ask questions when they perceive failures
- Granularity of feedback is vitaly important
- Users need transparency and access to knowledge provenance
- Users require explainable verification to build trust

Study also identified question types most important to users, to motivate future work

SPARK Wrapper and Database
Designed to

- Extract an explanation-relevant snapshot of execution state, including learning provenance
- Communicate that information to the Explainer

System utilizes introspective predicates to build a justification of current task(s), then outputs the justification structure via an RDBMS
- Database schema designed for easy, scalable storage and retrieval of relevant execution and provenance information
- Designed to be reusable for other types of reasoning about execution and action

Dialogue
Given a PML justification of SPARK’s execution, ICEE provides a dialogue interface to explaining actions. Users can start a dialogue with any of several question types, for example:

“Why are you doing <task>?”

ICEE contains several explanation strategies for each question type and, based on context and a user model, chooses one strategy, for example:

Strategy: Reveal task hierarchy
“i am trying to do <high-level-task> and <task> is one subgoal in the process.”

ICEE then parses the execution justification to present the portions relevant to the query and the explanation strategy. Strategies for explaining learned and modified procedures focus on provenance information and the learning method used. ICEE also suggests follow-up queries, enabling back-and-forth dialogue between agent and the user.

www.inference-web.org

TW funding from: DARPA, NSF, IARPA, ARL, Lockheed Martin, Fujitsu, SRI, IBM.