

# Explaining Cognitive Assistants that Learn

## CALO Year 3 ICEE Advances

Deborah McGuinness, Alyssa Glass\*  
 Knowledge Systems, Artificial Intelligence Lab  
 Computer Science Department, Stanford University  
**STANFORD UNIVERSITY**

Michael Wolverton, Alyssa Glass\*  
 SRI International  


### Motivation

Usable cognitive 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 that can supply interactive access to provenance, justifications, assumptions, and learned information.

We aim to improve trust in cognitive agents that learn by providing transparency concerning:

- Provenance
- Information manipulation
- Task processing
- Learning



#### New Work

- Explanation component for learning by instruction
- User Trust Study
- Design for explanation of preferences

### Strategies for Explaining Learning

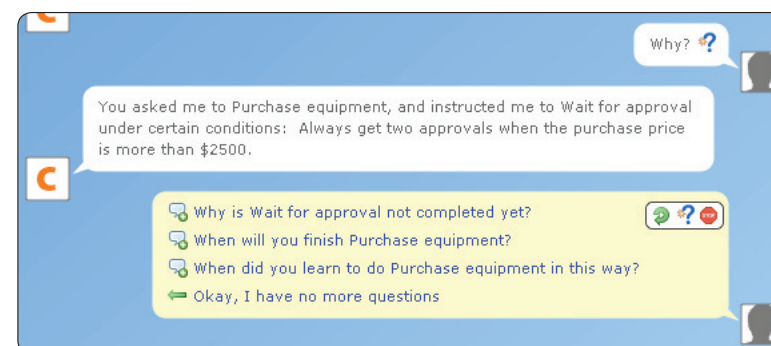
Given a PML justification of SPARK's execution, ICEE provides a dialogue infrastructure for explaining SPARK's actions, from within the existing Towel interface.

- ICEE contains a set of supported question types, and multiple explanation strategies for answering each type, based on context and a user model.
- Strategies for explaining learned and modified procedures focus on provenance information and the learning method used.
- Example learning-based strategies:
  - Reveal date & time of learning
  - Reveal data or other information used to learn
  - Reveal user-provided rationale for modification
  - Reveal type of modification
  - Reveal type of learning
- Initial focus on learning by instruction (Tailor)
- Expanding strategies to cover learning by demonstration (as in LAPDOG) and learning preferences using SVMs (as in PLIANT)



### 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 designed to explain cognitive agents, with focus on explaining task processing and learning



### User Trust Study

Interviewed 10 Critical Learning Period participants

- programmers, researchers, managers, and administrators
- wide range of ages and prior CALO experience

Focus of Study

- Trust
- Failures, surprises, and other sources of confusion
- Desired questions to ask CALO

Initial Results

- Explanations are required in order to trust agents that learn
- To build trust, users want transparency and provenance
- Identified question types most important to CALO users, to motivate future work



### Future Work

- Broaden explanation of learning and CALO integration
  - Explain learning by demonstration, integrating initially with CALO component LAPDOG
  - Explain preference learning, integrating initially with CALO component PTIME
- Investigate explanation of conflicts. Explore this as a driver to initiate learning procedure modifications or learning new procedures.
- Expand dialogue-based interaction and presentation of explanations, expanding our integration with Towel
- Use trust study results to prioritize provenance, strategy, and dialogue work.

### For More Information

McGuinness, D.L., Glass, A., Wolverton, M., and Pinheiro da Silva, P. Explaining Task Processing in Cognitive Assistants that Learn. Proceedings of the AAAI 2007 Spring Symposium on Interaction Challenges for Intelligent Assistants, March 2007. Available at: [www.ksl.stanford.edu/publications/](http://www.ksl.stanford.edu/publications/)

### Acknowledgements

Paulo Pinheiro da Silva was instrumental in early ICEE. We also thank Cynthia Chang and Li Ding for foundational Inference Web work; Jim Blythe for requirements and design for explanation of learning by instruction; Karen Myers, Ken Conley, David Morley, and Vasco Furtado for valuable discussions and feedback.

