Title

COmmon Ground Learning and Explanation (COGLE)

Team Composition

Palo Alto Research Center, Inc. (PARC), Mark Stefik, Principal Investigator, stefik@parc.com
Carnegie Mellon University, Christian Lebiere
Florida Institute for Human and Machine Cognition, Peter Pirolli
United States Military Academy, Robert Thomson

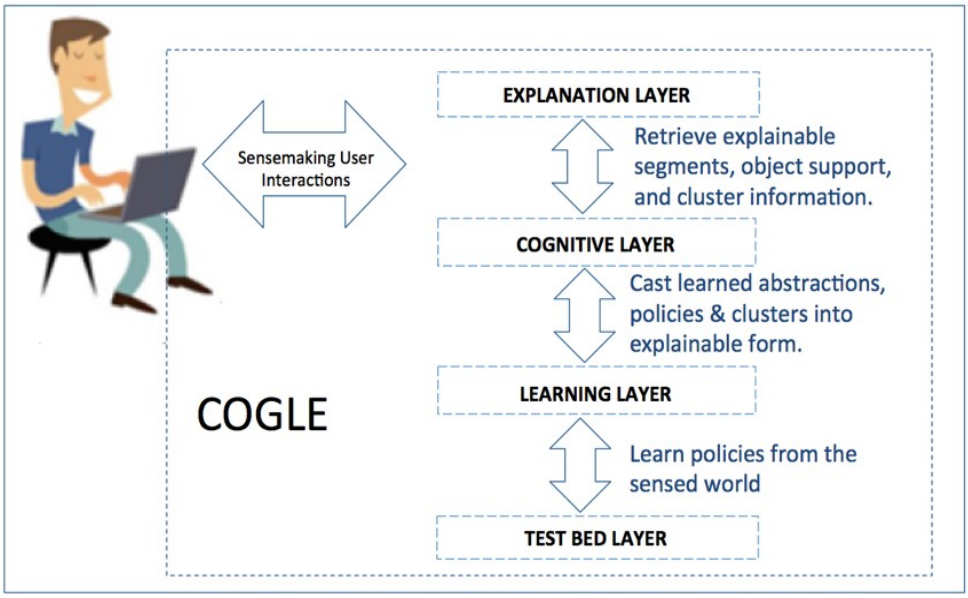
Problem Statement

PARC is developing an interactive sensemaking system called COGLE (COmmon Ground Learning and Explanation) for explaining the learned performance capabilities of an autonomous system; the representations, processes, and competencies underlying those capabilities; and the history that produced that learning. Explanations support rapid abstraction, composition, and generalization of causal patterns, and generalized causal patterns support rapid inference and prediction for novel specific cases.

Approach and Technique

COGLE (see Figure 1) uses a cognitive modeling layer to translate policies and performances of a machine-learned agent into information for constructing explanations. The program's Phase 1 challenge implicitly takes measuring user performance on a prediction task as a proxy for measuring user understanding of an XAI agent. Phase 1 user studies measured the degree to which explanations improve the accuracy of predicting the performance of autonomous systems. The COGLE team approached explanation as diagnostic support where explanations help study participants to understand an agent's competencies. The idea was that users need to understand which agent competencies are weak or strong in order to make their principled predictions about agent performance.

Figure 1. COGLE overview



Approved for Public Release, Distribution Unlimited

During Phase 1, competencies were defined as the knowledge/policy needed to solve a sub-goal. In this way, competencies were characterized as teleological. Different modules of a competency were created based on different training. These competency modules succeeded or failed under different conditions according to their training. The results of a user study are summarized below and a link to a full report is provided.

In Phase 2 the challenge problem requires study participants to identify an agent's high and low performing learned competencies. However, based on Phase 1 experience, teleological characterizations of competency are not precise enough because they do not account for the range of conditions that autonomous systems may face, especially in open worlds.

How should competencies be characterized? What do end user study participants need to understand about competencies? Should users be told what the possible competencies are? If an agent performs poorly, how does a study participant characterize the precise conditions under which the agent's competencies are weak?

In Phase 2, these questions are leading the team to pursue the development of a description language for characterizing competencies. Users describe the conditions under which competencies succeed or fail.

Figure 2 below illustrates an example interface view without explanation and Figure 3 below illustrates example explanation interfaces views.

Figure 2. Interface view without explanation

No Explanation

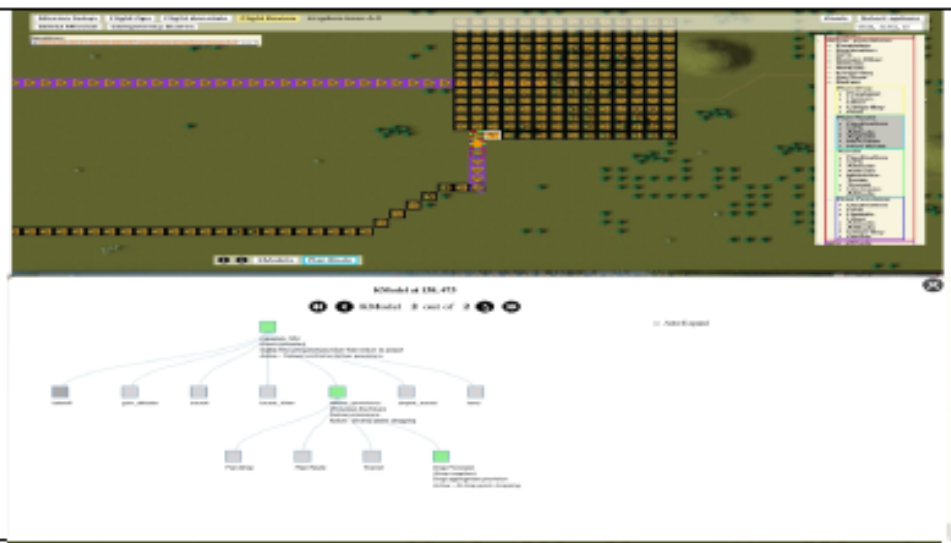
Showing UAS flight controls but no explanations.

Approved for Public Release, Distribution Unlimited

Figure 3. Explanation interface views

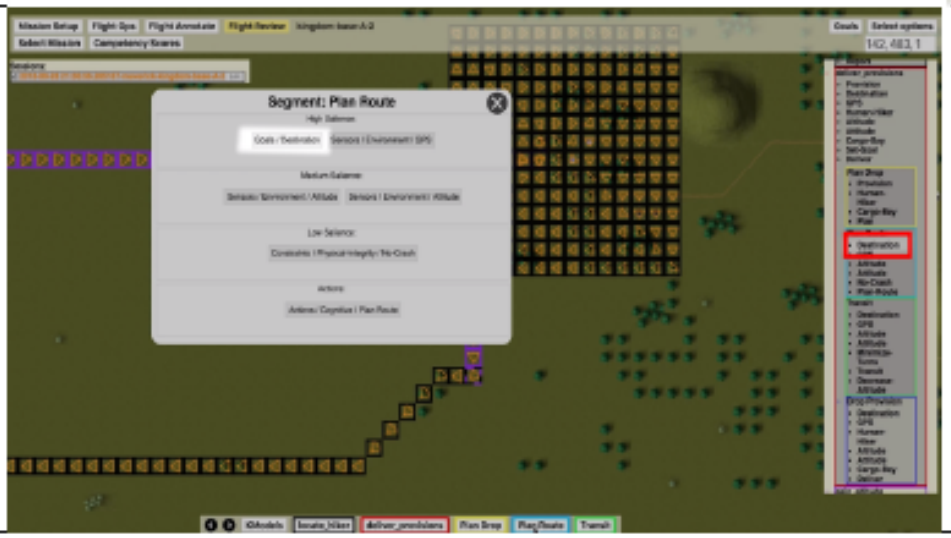
Explanation interface

Showing knowledge model, agent's active subgoals, and narrative interface



Explanation Interface

Showing saliences for what the agent is paying attention to.



Explanation Interface

Showing risk assessment and factored evaluation of performance of subgoals.

Choose Pilot: maverick

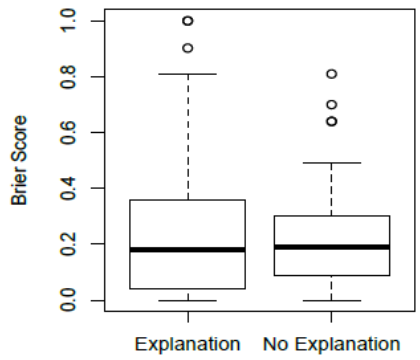
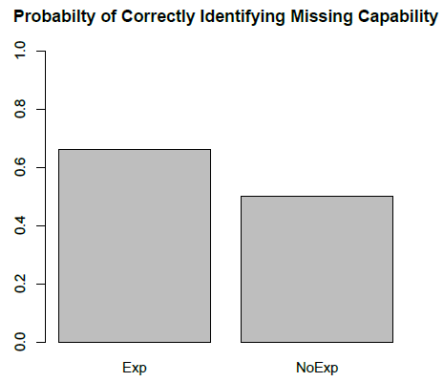
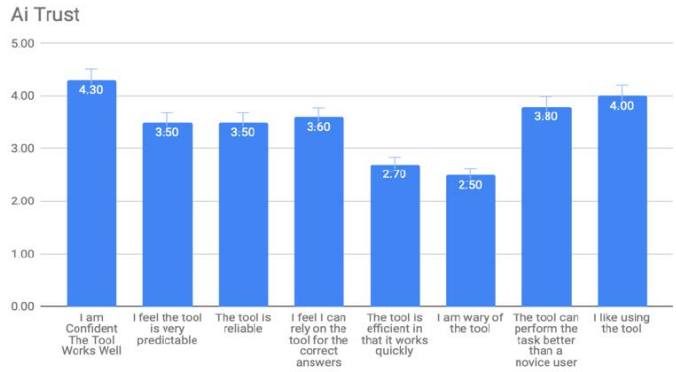
uccess	Failures	Failure Reason	Take-Off	Gain-Altitude	Transit	Locate-Hiker	Provision	Airport Transit	Landi
7	0		7.0	10.0	10.0	10.0	10.0	10.0	10.0
0	7	Goal provision(s) not successfully dropped to hiker	2.0	10.0	10.0	10.0	1.0	10.0	10.0
5	1	Goal provision(s) not successfully dropped to hiker. No attempt to land at goal airport	7.0	10.0	10.0	10.0	8.2	10.0	10.0
0	6	Goal provision(s) not successfully dropped to hiker	2.0	10.0	10.0	10.0	1.0	10.0	10.0

Approved for Public Release, Distribution Unlimited

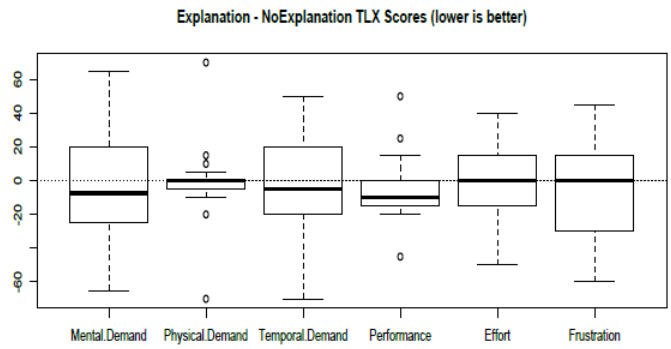
Key Objectives and Current Status

Table 1 below summarizes the COGLE measures, objectives, and actuals achieved to date.

Table 1. Measures, Objectives, and Actuals to Date

Measures	Objectives	Actuals to Date
Forecast Accuracy	Every participant makes a percent success forecast (0 - 100%) for a mission for each drone	
Identifying Missing Capability	Every participant identifies a (single) weak competency for an agent	
XAI Trust	Multi-factor measurement of study participant trust in the XAI system	

Approved for Public Release, Distribution Unlimited

NASA TLX* (perceived workload)	Multi-factor measurement of cognitive load	

*https://humansystems.arc.nasa.gov/groups/tlx/

	Size	Length	Type
User Study	18	2 hours per subject	Within subject

	Numerical Value		Hypothesis Results	Comments
	With Explanation	Without Explanation		
Improve forecasting predictions of agent success/ failure	0.2	0.2	No evidence; forecasting accuracies are all very high and there is a "floor effect"	**Brier Scores** $brier(pred, obs) = \frac{\sum(pred-obs)^2}{length(pred)}$ *Lower is better*

	Numerical Value		Hypothesis Results	Comments
	With Explanation	Without Explanation		
Improve confidence in predictions	---	---	Statistically marginal	Generalized linear model of the normalized gains

	Numerical Value		Hypothesis Results	Comments
	With Explanation	Without Explanation		
Yield an improved mental model of agent competencies	---	---	Statistically marginal	Analysis of hand-coded representations of submitted user explanations

	Numerical Value		Hypothesis Results	Comments
	With Explanation	Without Explanation		
Not incur greater cognitive load	---	---	No statistical difference in cognitive load	NASA TLX scores

	Numerical Value		Hypothesis Results	Comments
	With Explanation	Without Explanation		
Yield appropriate trust	---	---	Yes	Responses to trust in AI questions

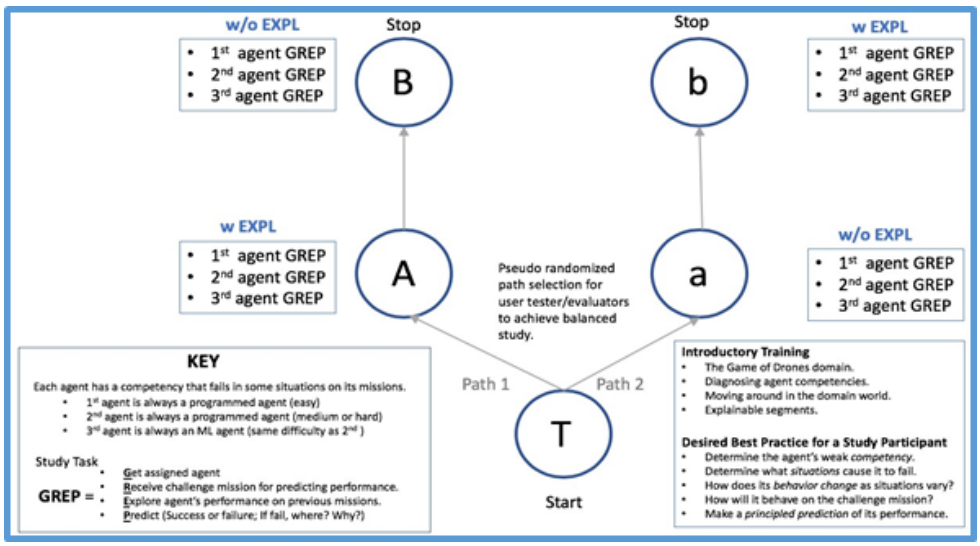
Approved for Public Release, Distribution Unlimited

	Numerical Value		Hypothesis Results	Comments
	With Explanation	Without Explanation		
Be the preferred interface	14	4	Significantly preferred	Binomial Test, $p < .05$

Test and Evaluation

An experiment (see Figure 4) that compared the performance of study participants on the Phase 1 challenge problem was conducted. Participants were asked to predict agent performance on a new mission given experience with the agent (with and without explanation interface) on two previous missions. Results from the study showed a "floor effect", meaning that participants did very well in predicting performance even without the explanation interface. The explanation interface did yield appropriate trust and was the significantly preferred interface.

Figure 4. Experiment Overview



It is believed that the floor effect seen in the Phase 1 experiments was due in part to a "dumbing down" of the test problems in the study. In hindsight, the test problems were made simple for two reasons. On the basis of early trials, it was believed that users had difficulty understanding more complex UAS missions. To this end, test problems were designed where only one competency was weak (single failure assumption). On simple domain problems, this was too easy for users.! Secondly, the reinforcement learning algorithms that were deployed in Phase 1 proved could not solve complex domain tasks.

Pirolli, P., Schooler, J. *Exigisi Evaluation 2018*, (Detailed User Study Evaluation Report for XAI Evaluation team at the Naval Research Laboratory), https://bit.ly/2YYmkhk.

For additional information:

Stefik, M. *COGLE Demonstration Video*, February 2019, https://bit.ly/2OP1tsb.

Somers, S., Mitsopoulos, K., Lebiere, C., and Thomson, R. *Cognitive-Level Salience for Explainable Artificial Intelligence.* Submitted to the 41st Annual Meeting of the Cognitive Science Society. Montreal, CA. July 24-27, 2019.

Approved for Public Release, Distribution Unlimited

Somers, S., Mitsopoulos, K., Lebiere, C., and Thomson, R. *Explaining Decisions of a Deep Reinforcement Learner with a Cognitive Architecture.* In Proceedings of the 16th Annual Meeting of the International Conference on Cognitive Modeling. Madison, WI. July 21-24, 2018.

Stefik, M., and G. Michael Youngblood, *The Who, What, When, and Why of Explainable AI*. COGLE Working Paper. May 2, 2018. https://bit.ly/2KxrwBu

Stefik, M., and G. Michael Youngblood, *Understanding the Power of Goals and Abstractions to Guide Explanation and Cognition*. COGLE Working Paper. May 2, 2018. https://bit.ly/2HNs2xp

Stefik, M. and G. Michael Youngblood, *An Introspective Discourse Model for Creating Common Ground with XAI Systems*. COGLE Working Paper. May 2, 2018. https://bit.ly/2KzkUCN

Zhu, J., Liapis, A., Risi, S., Bidarra, R., and Youngblood, G. M. (2018, August). *Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation.* In 2018 IEEE Conference on Computational Intelligence and Games (CIG) (pp. 1-8). IEEE.

Approved for Public Release, Distribution Unlimited