Title

Learning and Communicating Explainable Representations for Analytics and Autonomy

Team Composition

University of California, Los Angeles (UCLA), Song-Chun Zhu, Principal Investigator, sczhu@stat.ucla.edu
UCLA, Yingnian Wu
Oregon State University, Sinisa Todorovic
Michigan State University, Joyce Chai

Problem Statement

The UCLA team is developing an explainable AI framework for two domains: (a) multi-model analytics for complex event understanding, and (b) autonomy by recognition, reasoning, and planning. For these two domains, the work can be broadly divided into three main directions aimed at developing:

> **A Performer** (i.e., a vision or robot system) for complex settings, in which it's predictions, plans, or decisions are not readily obvious to the user from one-shot explanations, but require non-trivial sequences of explanations (i.e. a dialogue with the user);
>
> **An Explainer** that generates maximum utility explanations in a dialog with the user; and
>
> **An Evaluator** that quantitatively evaluates the effect of explanations on the user's understanding of the machine's mental model.

Approach and Technique

UCLA's explainable AI framework is based on Theory of Mind (ToM). ToM is the ability to reason about the perceptions and beliefs of others, in addition to one's own perceptions and beliefs. The framework, X-ToM, generates interactive and collaborative explanations by incorporating the machine's understanding of the user's mind; and evaluates the user's trust and reliance in the machine by explicitly measuring the user's understanding of the machine's mind. To facilitate this collaboration, X-ToM explicitly models the mental states of both machine and user visual understanding using the following three parse graphs (pg) in the form of an And-Or Graph (AOG):

> pg^{M}: the machine's own inference about objects and their locations in the image/video;
>
> pg^{UinM}: the user's mind as inferred by the machine; and
>
> pg^{MinU}: the machine's mind as inferred by the user.

UCLA's approach has two important phases: In the first phase, they use reinforcement learning to train the Explainer to provide optimal explanations by inferring and updating pg^{UinM} at every turn in the dialog. Optimal explanations generated by the Explainer are the key to maximize the user's trust in the machine. In the second phase, the Evaluator assesses the extent to which the explanations provided in the first phase have helped the user understand the machine's behaviors. Specifically, the user is asked to solve new tasks similar to (but different from) the ones used in the first phase. Based on the user predictions, the Evaluator estimates pg^{MinU} and quantifies user trust and reliance in the machine by comparing pg^{MinU} and pg^{M}. Figure 1 depicts X-ToM for optimizing the dialog with a user towards estimating and increasing user trust.

Approved for Public Release, Distribution Unlimited

Figure 1. X-ToM for optimizing dialog

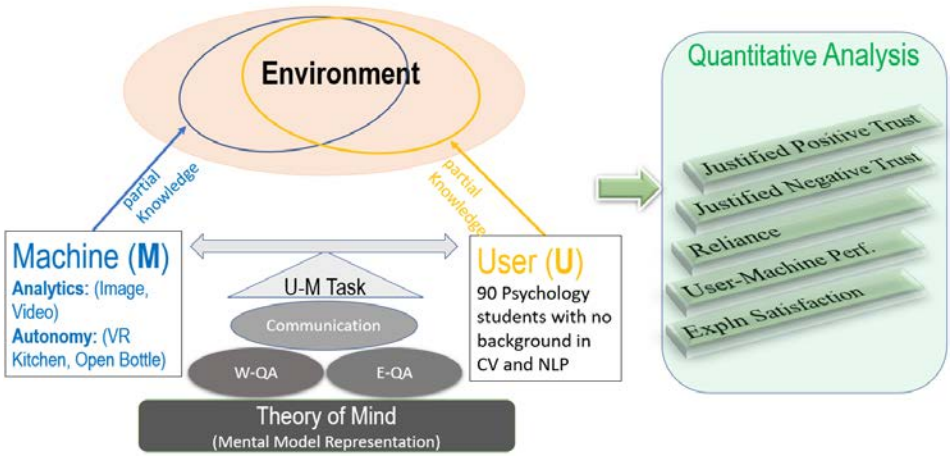
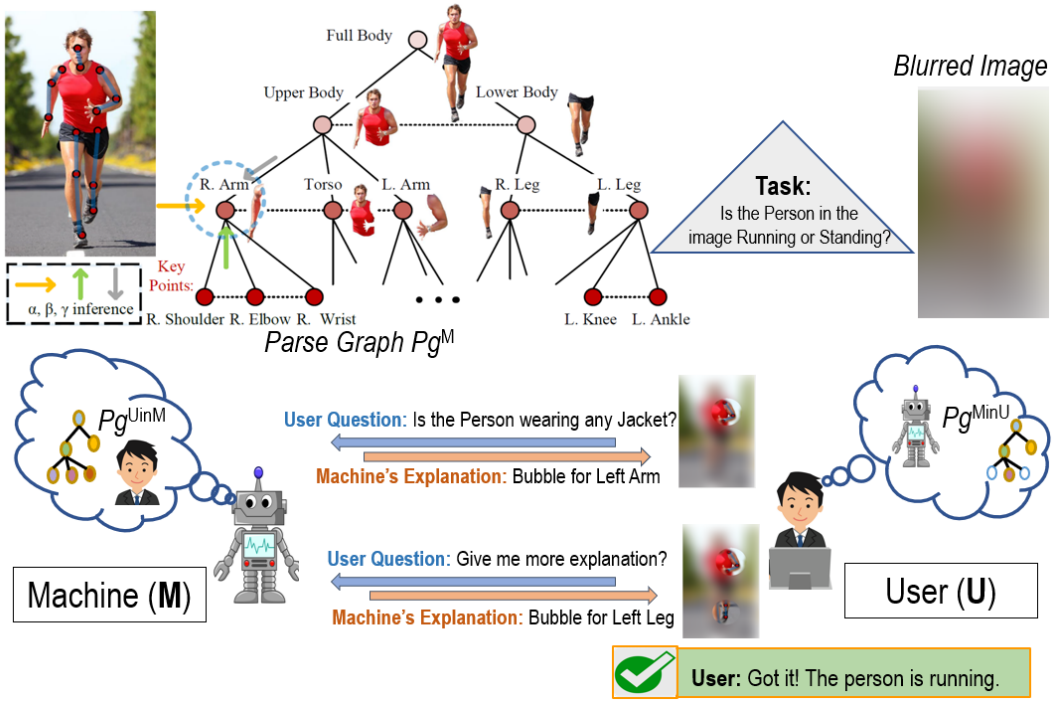


Figure 2 illustrates an example of collaborative task solving in image analytics where trust and reliance in, terms of pg^{M}, pg^{UinM}, and pg^{MinU}, are estimated through dialog.

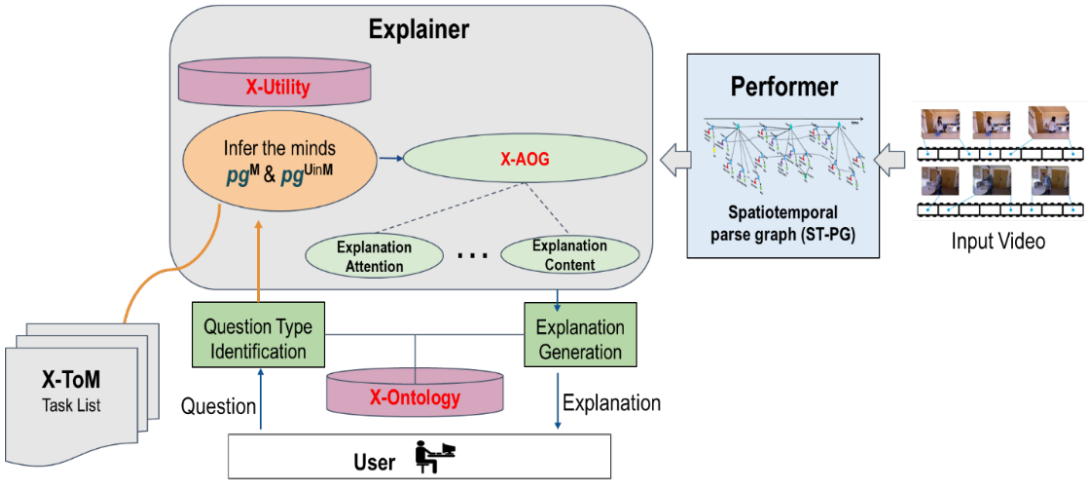
Figure 2. Collaborative task solving



Approved for Public Release, Distribution Unlimited

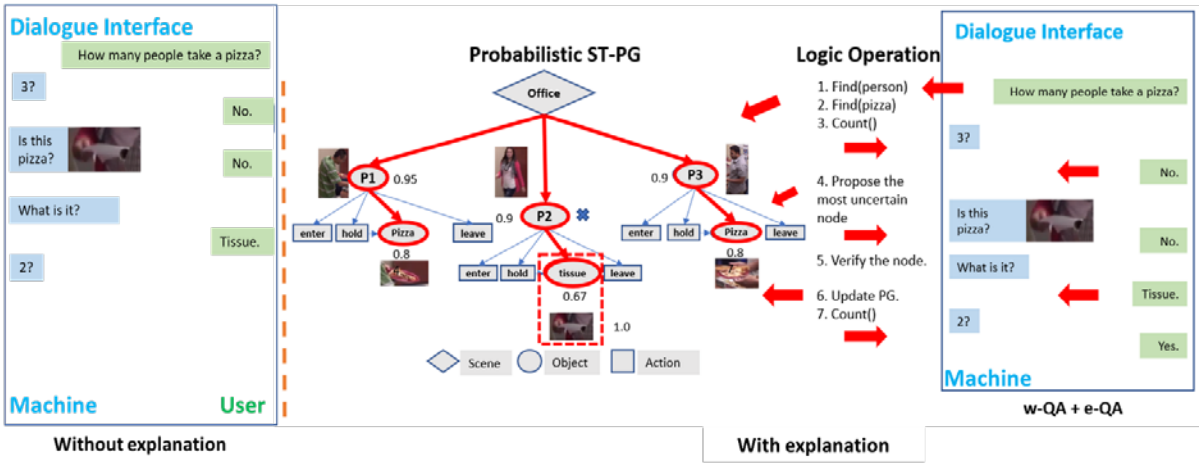
UCLA's integrated analytics system enables many functions in a joint spatial, temporal, and causal parsing framework. As shown in Figure 3, the Performer can obtain a scene-centric PG, with spatio-temporal graphs (ST-PG) for each person, generated by different vision and cognition algorithms. Then, the Explainer infers the user's mind via the X-Utility module in terms of the user question, and further generates the explanation via the X-AOG module. The X-Ontology module contributes to the explanation generation.

Figure 3. X-ToM for video analytics



UCLA's dialogue interface is shown in Figure 4, illustrating both without explanation and with explanation. The X-ToM system focuses on generating a uniform and consistent explanation from the probabilistic ST-PG to the logic operation according to the user's question. Specifically, user questions are first transformed into logical operations. For example, "How many people will take a pizza?" can be translated into: "1. Find(person); 2. Find(pizza), and; 3. Count()". Next, the logical operations are performed on the probabilistic ST-PG. In this way, the user can comprehensively understand the machine to improve the pg^{M}, while the machine can estimate the trust and reliance of the user by predicting the pg^{UinM} and pg^{MinU}.

Figure 4. Dialogue interface without and with explanation

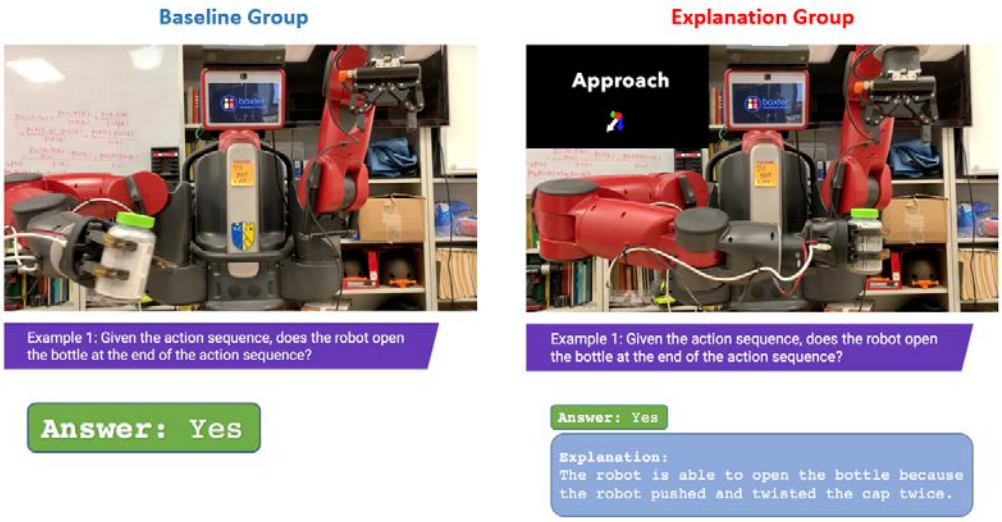


A user-machine task was designed to measure user trust at certain stages of robotic execution, and evaluate how user trust changes over the course of execution. The objective of the study was to

Approved for Public Release, Distribution Unlimited

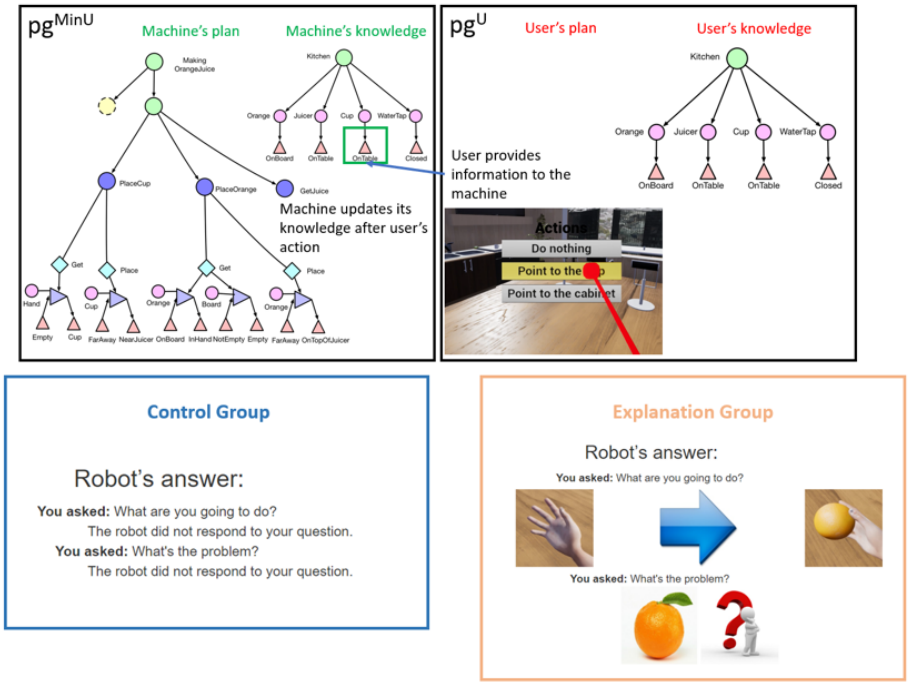
determine which stages of robotic execution are most critical in influencing user trust; and what information is most effective in promoting user trust. Figure 5 illustrates an example of the user-machine task without explanation (baseline group, left) and with explanation (explanation group, right).

Figure 5. X-ToM for a force-based robot grasping system



ULCA has also designed studies in which both the machine and user solve common household tasks in a virtual reality kitchen. They have implemented manipulations such as cutting, pealing, boiling, grading, grasping, and opening; as well as the corresponding fluent changes for relevant objects (including food ingredients, kitchenware, and other tools) within the virtual reality kitchen. Figure 6 illustrates an example where the machine provides explanations to the user based on its knowledge and inference of the user's mind.

Figure 6. X-ToM for Human Robot Teaming



Approved for Public Release, Distribution Unlimited

Key Objectives and Current Status

Tables 1 - 4 below summarize UCLA's measures and actual results to date.

Table 1. Image analytics results

	Size	Length	Type
User Study	120 participants from UCLA Psychology Subject Pool	3 conditions (without explanations, with attention map explanations, and with x-ToM explanations)	Between group study

	Numerical Value			Hypothesis Results	Measure Definition
	Without Explanation	**Attention Maps as Explanation**	**X-ToM Explanation**		
Justified Positive Trust for Detection (JPT_Detection)	9.1%	25.3%	84.1%	Significant, $p < 0.01$	in detecting and localizing objects and parts of interest or a user activity appearing in the image
Justified Negative Trust for Detection (JNT_Detection)	32.4%	28.1%	74.0%	Significant, $p < 0.01$	in detecting and localizing objects and parts of interest or a user activity appearing in the image
Justified Positive Trust for Influence (JPT_Influence)	24.0%	35.8%	76.4%	Significant, $p < 0.01$	for identifying the most influential factors in machine's successful or failed detection of objects or parts of interest, or a user activity appearing in the image
Justified Negative Trust for Influence (JNT_Influence)	4.3%	17.9%	84.2%	Significant, $p < 0.01$	for identifying the most influential factors in machine's successful or failed detection of objects or parts of interest, or a user activity appearing in the image

	Numerical Value			Hypothesis Results	Comments
	Without Explanation	**Attention Maps as Explanation**	**X-ToM Explanation**		
User-Machine Performance/Prediction Accuracy	31.1%	42.0%	85.7%	Significant, $p < 0.01$	User-machine performance is significantly higher in X-ToM ($p < 0.01$)

Approved for Public Release, Distribution Unlimited

					compared to attention maps and question answer baselines User-machine performance of attention map is not significantly higher than question answer baselines

	Numerical Value			Hypothesis Results	Comments
	Without Explanation	**Attention Maps as Explanation**	**X-ToM Explanation**		
Response Times	16 second	20 seconds	17 seconds	Not significant	Record time taken by the user in solving a task (or answering a question) in the testing stage Results show no significant difference in the response times across the three groups

	Numerical Value			Hypothesis Results	Measure
	Without Explanation	**Attention Maps as Explanation**	**X-ToM Explanation**		
Explanation Satisfaction	4.20	4.05	7.81	Significant, $p < 0.01$	Usefulness (Likert scale 0 - 9)
	4.36	3.68	7.22	Significant, $p < 0.01$	Sufficiency (Likert scale 0 - 9)
	3.18	3.40	7.70	Significant, $p < 0.01$	Appropriate detail (Likert scale 0 - 9)
	5.80	5.42	6.33	Not significant	Confidence (Likert scale 0 - 9)
	6.60	7.31	8.04	Not significant	Understandability (Likert scale 0 - 9)
	6.00	5.43	6.80	Not significant	Accuracy (Likert scale 0 - 9)
	6.10	5.80	7.21	Not significant	Consistency (Likert scale 0 - 9)

	Numerical Value			Hypothesis Results	Comments
	Without Explanation	**Attention Maps as Explanation**	**X-ToM Explanation**		
Qualitative Reliance	5.4	5.1	7.6	Significant, $p < 0.01$	Likert scale of 0 – 9; qualitative

Approved for Public Release, Distribution Unlimited

					reliance values collected from users is consistent with X-ToM quantitative reliance values; users' qualitative reliance in attention maps is lower compared to the question answering baseline

Table 2. Video analytics results

	Numerical Value		Hypothesis Results
	Without Explanation	**X-ToM Explanation**	
Justified Positive Trust	41.0%	93.2%	Significant, $p < 0.01$
Justified Negative Trust	24.3%	54.5%	Not significant
Reliance	41.7%	72.1%	Significant, $p < 0.01$

Table 3. Robotic "open bottle" results

	Numerical Value		Hypothesis Results	Comments
	Without Explanation	**X-ToM Explanation**		
Justified Positive Trust (JPT)	39.1%	64.7%	Significant, $p < 0.01$	Likert scale of 1-7. To predict bottle opening success
Justified Negative Trust (JNT)	54.4%	46.5%	Significant, $p < 0.01$	Likert scale of 1-7. To predict bottle opening failure

Table 4. Virtual reality kitchen results

	Numerical Value		Hypothesis Results	Comments
	Without Explanation	**X-ToM Explanation**		
Prediction Accuracy	12.1%	77.4%	Significant, $p < 0.01$	N/A
Explanation Satisfaction	16.1%	24.5%	Not Significant	Efficiency; Likert scale of 0 - 5
	24.3%	36.0%	Not Significant	Appropriate detail; Likert scale of 0 - 5
	25.0%	38.3%	Not Significant	Usefulness; Likert scale of 0 - 5

Approved for Public Release, Distribution Unlimited

	23.6%	37.4%	Not Significant	Clarity; Likert scale of 0 - 5
User-Machine Performance	40.0%	69.4%	Significant, $p < 0.01$	N/A

Test and Evaluation

UCLA conducted user studies to assess the effectiveness of the X-ToM Explainer, trained on Amazon Mechanical Turk, in increasing user trust through explanations. 120 participants were selected from the UCLA Psychology Department student pool. The participants had no background in computer vision, deep learning, or natural language processing. Each participant went through an introduction phase where they were introduced to the tasks. Then they went through a familiarization phase, where they became familiar with the machine's underlying inference processes (Performer), followed by a testing phase where trust metrics were used to assess their trust in the Performer.

Arjun Akula, Changsong Liu, Sari Sadiya, Sinisa Todorovic, Joyce Chai, Song-Chun Zhu. Gaining Justified Trust by Explaining with Theory of Mind. In International Joint Conference on Artificial Intelligence (IJCAI 2019) [submitted].

Xiaofeng Gao, Ran Gong, Tianmin Shu, Xu Xie, Shu Wang, Song-Chun Zhu. VRKitchen: an Interactive 3D Virtual Environment for Task-oriented Learning. arXiv:1903.05757v1, Mar 2019.

W. Wang, Y. Xu, Q.S. Zhang, J. Shen, S.-C. Zhu. Deep Structured Network with Joint and Interpretable Bottom-up and Top-down Inference. AAAI Conference on Artificial Intelligence (AAAI), 2019 (under review).

S. Yang, Q. Gao, S. Saba-Sadiya, and J. Y. Chai. Commonsense Justification for Action Explanation. Conference on Empirical Methods in Natural Language Processing (EMNLP), Brussels, Belgium, November 1-4, 2018.

Q.S. Zhang, R. Cao, F. Shi, Y.N. Wu, and S.-C. Zhu. Interpreting CNN Knowledge Via An Explanatory Graph. AAAI Conference on Artificial Intelligence (AAAI), 2018

Q.S. Zhang, W.G. Wang, and S.-C. Zhu. Examining CNN Representations With Respect To Dataset Bias. AAAI Conference on Artificial Intelligence (AAAI), 2018

J. Y. Chai, Q. Gao, L. She, S. Yang, S. Saba-Sadiya, and G. Xu. Language to Action: Towards Interactive Task Learning with Physical Agents. Proceedings of the 27th International Joint Conference on Artificial Intelligence and 23rd European Conference on Artificial Intelligence (IJCAI-ECAI), Stockholm, Sweden, July 13-19, 2018.

Q. Gao, S. Yang, J. Y. Chai, and L. Vanderwende. What Action Causes This? Towards Naïve Physical Action-Effect Prediction. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, Australia, July 15-20, 2018.

J. Y. Chai, M. Cakmak, and C. Sidner. Teaching Robots New Tasks through Natural Interaction. In: Interactive Task Learning: Agents, Robots, and Users Acquiring New Tasks through Natural Interactions, ed. K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26, J. Lupp, series editor. Cambridge, MA: MIT Press. [In Press]

A. L. Thomas, E. Lieven, M. Cakmak, J. Y. Chai, S. Garrod, W. Gray, S. Levinson, A. Paiva, N. Russwinkel. Interaction for Task Instruction and Learning. In Gluck, K. A. and Laird, J. E. eds., Interactive Task

Approved for Public Release, Distribution Unlimited

Learning: Agents, Robots, and Users Acquiring New Tasks through Natural Interactions. Strüngmann Forum Reports, J. Lupp, series editor, volume 26, MIT Press [In Press].

Q. Gao, L. She, and J. Y. Chai. Interactive Learning of State Representation through Natural Language Instruction and Explanation. AAAI 2017 Fall Symposium on Natural Communication for User-Robot Collaboration.

Q. Zhang, Y. N. Wu and S.-C. Zhu, Interpretable Convolutional Neural Networks. Computer Vision and Pattern Recognition (CVPR), 2018.

A. Roy and S. Todorovic, "Learning to learn second-order back-propagation for CNNs using LSTMs," in Proc. 24rd International Conference on Pattern Recognition (ICPR), Beijing, China, 2018.

P. Lei, F. Li, and S. Todorovic, "Boundary Flow: A Siamese network that predicts boundary motion without training on motion," in Proc. IEEE Comp. Soc. Conf. Computer Vision Pattern Recognition (CVPR 2018), Salt Lake City, UT, 2018.

P. Lei and S. Todorovic, "Temporal Deformable Residual Networks for action segmentation in videos," in Proc. IEEE Comp. Soc. Conf. Computer Vision Pattern Recognition (CVPR 2018), Salt Lake City, UT, 2018.

Y. Xu, E. Markowitz, T. Liu, W. Wang, J. Xie, S.-C. Zhu. Learning 3D User Shapes, Poses and Actions from 3D Animations. AAAI Conference on Artificial Intelligence (AAAI), 2019 (under review).

T. Liu, X. Xie, Y. Xu, W. Wang, S.-C. Zhu. Video Game Exploit: Hacking Game Assets to Learn 3D User-Object Interactions. AAAI Conference on Artificial Intelligence (AAAI), 2019 (under review).

M. Edmonds, J. Kubricht, C. Summers, Y. Zhu, B. Rothrock, S.C. Zhu, & H. Lu. User Causal Transfer: Challenges for Deep Reinforcement Learning. 40th Annual Meeting of the Cognitive Science Society (CogSci), 2018.

For additional information:

X-ToM Illustration Video: https://www.youtube.com/watch?v=cwnOXTa1DAU

X-ToM Policy Learning Video: https://www.youtube.com/watch?v=EqBM8iN-AMg

X-ToM Introduction Phase Video: https://www.youtube.com/watch?v=dtmXLxB1yV8

X-ToM Familiarization Phase Video: https://www.youtube.com/watch?v=6i5KJmqPSHY

X-ToM Testing Phase Video: https://www.youtube.com/watch?v=LOFKcqPqfJs

X-ToM for video analytics: http://xai.visualturingtest.org/

Approved for Public Release, Distribution Unlimited