Visualizations for an Explainable Planning Agent

Tathagata Chakraborti[1], Kshitij P. Fadnis[2], Kartik Talamadupula[2], Mishal Dholakia[2]
Biplav Srivastava[2], Jeffrey O. Kephart[2], Rachel K. E. Bellamy[2]
[1] Computer Science Department, Arizona State University, Tempe, AZ 85281 USA
[2]IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA
tchakra2@asu.edu, { kpfadnis, krtalamad, mdholak, biplavs, kephart, rachel } @ us.ibm.com

Abstract

In this demonstration, we report on the visualization capabilities of an Explainable AI Planning (XAIP) agent that can support human in the loop decision making. Imposing transparency and explainability requirements on such agents is crucial for establishing human trust and common ground with an end-to-end automated planning system. Visualizing the agent's internal decision making processes is a crucial step towards achieving this. This may include externalizing the "brain" of the agent: starting from its sensory inputs, to progressively higher order decisions made by it in order to drive its planning components. We demonstrate these functionalities in the context of a smart assistant in the Cognitive Environments Laboratory at IBM's T.J. Watson Research Center.

Recently, there have been concerted efforts towards making the outputs of planning processes more palatable to human decision makers – e.g. eXplainable AI Planning (XAIP) [Fox *et al.*, 2017; Langley *et al.*, 2017]. One of the key features that an XAIP agent must support is visualization. For an end-to-end planning system – which goes from lower level sensory data (e.g. vision, speech) to progressively higher level decision-making capabilities (planning, plan recognition) – this becomes even more challenging. It is in this spirit that we present `Mr.Jones`, a set of visualization capabilities for an XAIP agent that assists with human-in-the-loop decision-making in an instrumented meeting space.

Introducing `Mr.Jones` – `Mr.Jones` [Chakraborti *et al.*, 2017c], situated in the `CEL` – the Cognitive Environments Laboratory – at IBM's T.J. Watson Research Center is designed to embody the key properties of a proactive assistant while fulfilling the properties desired of an XAIP agent. Similar to [Manikonda *et al.*, 2017], we divide the responsibilities of `Mr.Jones` into two processes (c.f. Figure 1) –- ***Engage***, where plan recognition techniques are used to identify the task in progress; and ***Orchestrate***, which involves active participation in the decision-making process via real-time plan generation, visualization, and monitoring.

Mind of `Mr.Jones` – The externalization of the "mind" of `Mr.Jones` – i.e. the various processes that feed the different capabilities of the agent (c.f. Figure 2) – consists of five widgets. The largest widget on the top represents the probability distribution that indicates the confidence of `Mr.Jones` in identifying the task being collaborated on, along with a button that displays the provenance of each such belief. The information used as provenance is generated directly from the plans used internally by the recognition module [Ramirez and Geffner, 2010] and justifies why, given the model of the underlying planning problems, these tasks look likely in terms of plans that achieve those tasks. The system is adept at handling uncertainty in its inputs (it is interesting to note that in coming up with an explanatory plan, it has announced likely assignments to unknown agents in its space).

Below this is a set of four widgets, each of which give users a peek into an internal component of `Mr.Jones`. The first of them (top left) presents a wordcloud representation of `Mr.Jones`'s belief in each of the tasks; the size of the word representing that task corresponds to the probability associated with that task. The second widget (top right) shows the agents that are recognized as being in the environment currently – this information is used by the system to determine what kind of task is more likely. This information is obtained from four independent camera feeds that give `Mr.Jones` an omnispective view of the environment; this information is represented via snapshots (sampled at 10-20 Hz) in the third widget (bottom left). Finally, the fourth widget (bottom right) represents a wordcloud based summarization of the audio transcript of the environment. This transcript provides a succinct representation of the things that have been said in the environment in the recent past via the audio channels. The interface thus provides a (constantly updating) snapshot of the various sensory and cognitive organs associated with `Mr.Jones` – the eyes, ears, and mind of the `CEL`. This is organized at increasing levels of abstraction –

[1] *Raw Inputs* – These show the camera feeds and voice capture (speech to text outputs) as received by the system. These help in externalizing what information the system is working with at any point of time, and can be used in debugging at the input level if the system makes a mistake or in determining whether it is receiv-

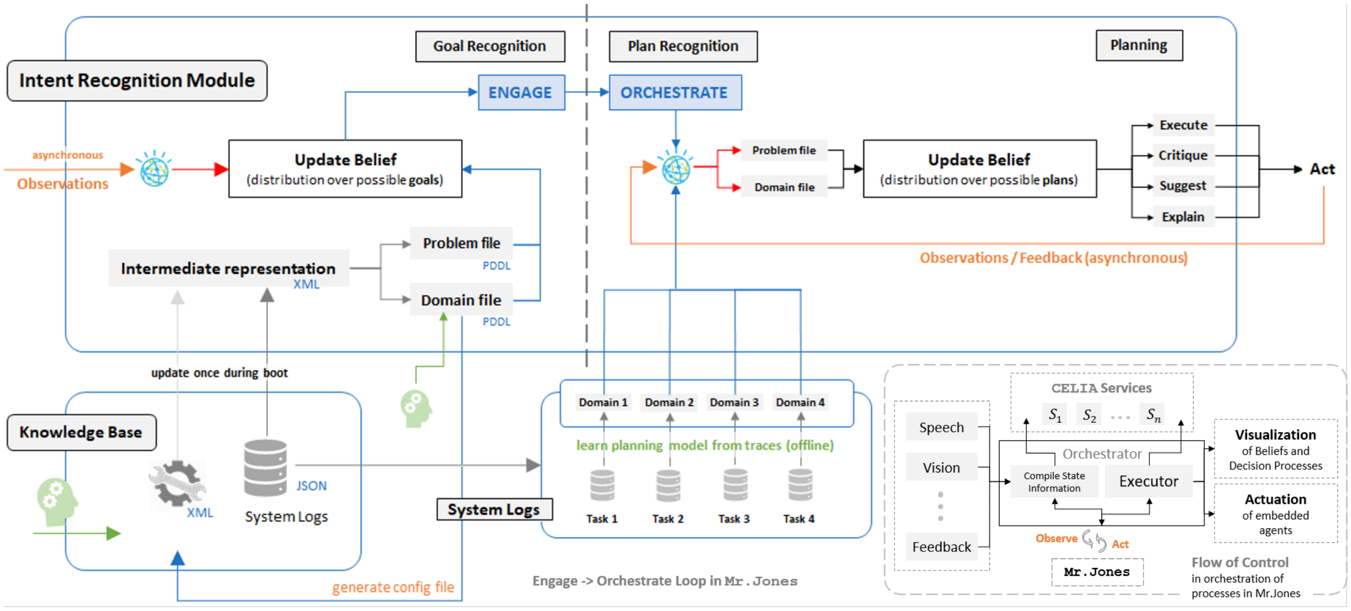

Figure 1: The building blocks of Mr.Jones – the *Engage* and *Orchestrate* processes situate it proactively in a decision support setting.

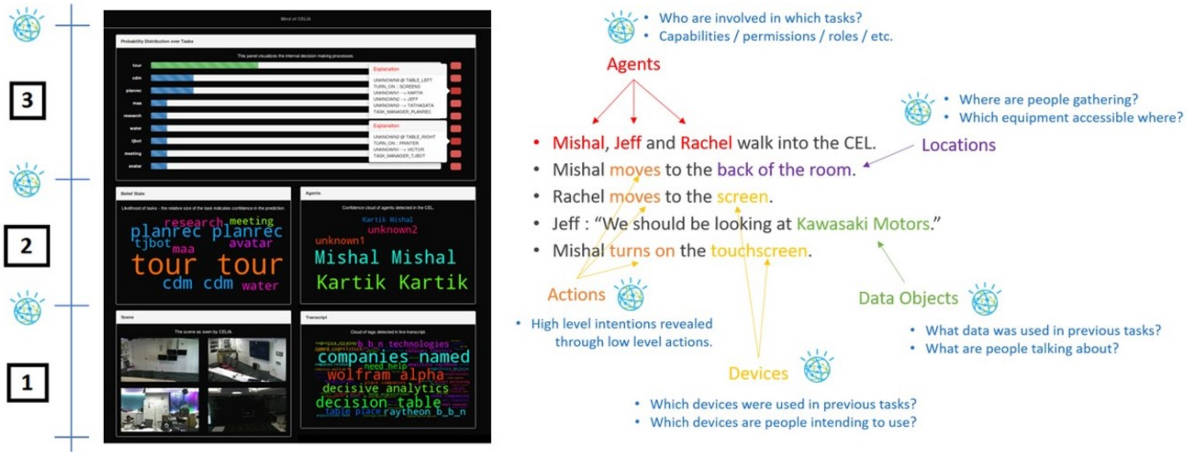
3
2
1
tour tour
cdm cdm water
planrec planrec
research meeting
Mishal Mishal
Kartik Kartik
unknown1
unknown2
companies named
wolfram alpha
decisive analytics
decision table
Agents
• Who are involved in which tasks?
• Capabilities / permissions / roles / etc.
• Mishal, Jeff and Rachel walk into the CEL.
• Mishal moves to the back of the room.
• Rachel moves to the screen.
• Jeff : "We should be looking at Kawasaki Motors."
• Mishal turns on the touchscreen.
Locations
• Where are people gathering?
• Which equipment accessible where?
Actions
• High level intentions revealed through low level actions.
Data Objects
• What data was used in previous tasks?
• What are people talking about?
Devices
• Which devices were used in previous tasks?
• Which devices are people intending to use?

Figure 2: Snapshot of the mind of Mr.Jones. A video of the system in action can be viewed at https://youtu.be/ZEHxCKodEGs.

ing enough information to make the right decisions. It is especially useful for an agent like Mr.Jones, which is not embodied in a single robot or interface but is part of the environment as a whole; in such cases, it is difficult to attribute specific events and outcomes to the agent.

[2] *Lower level reasoning* – The next layer deals with the first stage of reasoning over the raw inputs – *What are the topics being talked about? Who are the agents in the room? Where are they situated?* This helps a user identify what knowledge is being extracted from the input layer and fed into the reasoning engines. It increases the situational awareness of agents by visually summarizing the contents of the scene at any point of time.

[3] *Higher level reasoning* – Finally, the top layer uses information extracted at the lower levels to reason about abstract tasks in the scene. It visualizes the outcome of the plan recognition process, along with the provenance of the information extracted from the lower levels. This layer puts into context the agent's current understanding of the processes in the scene.

In addition to this, we support a plan visualization tool Fresco [Chakraborti *et al.*, 2017a], that builds on recent work in top-K planning [Katz *et al.*, 2018] and model-based plan explanations [Chakraborti *et al.*, 2017b] to provide a concise visualization of a plan. A detailed description of the system can be accessed at https://arxiv.org/abs/1709.04517.

Acknowledgments

A significant part of this work was completed while Tathagata Chakraborti was an intern at IBM's T. J. Watson Research Center. The continuation of his work at Arizona State University is supported by an IBM Ph.D. Fellowship.

References

[Chakraborti *et al.*, 2017a] Tathagata Chakraborti, Kshitij P. Fadnis, Kartik Talamadupula, Mishal Dholakia, Biplav Srivastava, Jeffrey O. Kephart, and Rachel K. E. Bellamy. Visualizations for an explainable planning agent. *CoRR*, abs/1709.04517, 2017.

[Chakraborti *et al.*, 2017b] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In *IJCAI*, 2017.

[Chakraborti *et al.*, 2017c] Tathagata Chakraborti, Kartik Talamadupula, Mishal Dholakia, Biplav Srivastava, Jeffrey O Kephart, and Rachel KE Bellamy. Mr. Jones – Towards a Proactive Smart Room Orchestrator. *AAAI Fall Symposium on Human-Agent Groups*, 2017.

[Fox *et al.*, 2017] Maria Fox, Derek Long, and Daniele Magazzeni. Explainable Planning. In *First IJCAI Workshop on Explainable AI (XAI)*, 2017.

[Katz *et al.*, 2018] Michael Katz, Shirin Sohrabi, Octavian Udrea, and Dominik Winterer. A Novel Iterative Approach to Top-k Planning. In *International Conference on Automated Planning and Scheduling (ICAPS)*, 2018.

[Langley *et al.*, 2017] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. Explainable Agency for Intelligent Autonomous Systems. In *AAAI/IAAI*, 2017.

[Manikonda *et al.*, 2017] Lydia Manikonda, Tathagata Chakraborti, Kartik Talamadupula, and Subbarao Kambhampati. Herding the crowd: Using automated planning for better crowdsourced planning. *Journal of Human Computation*, 2017.

[Ramirez and Geffner, 2010] Miquel Ramirez and Hector Geffner. Probabilistic plan recognition using off-the-shelf classical planners. In *AAAI*, 2010.