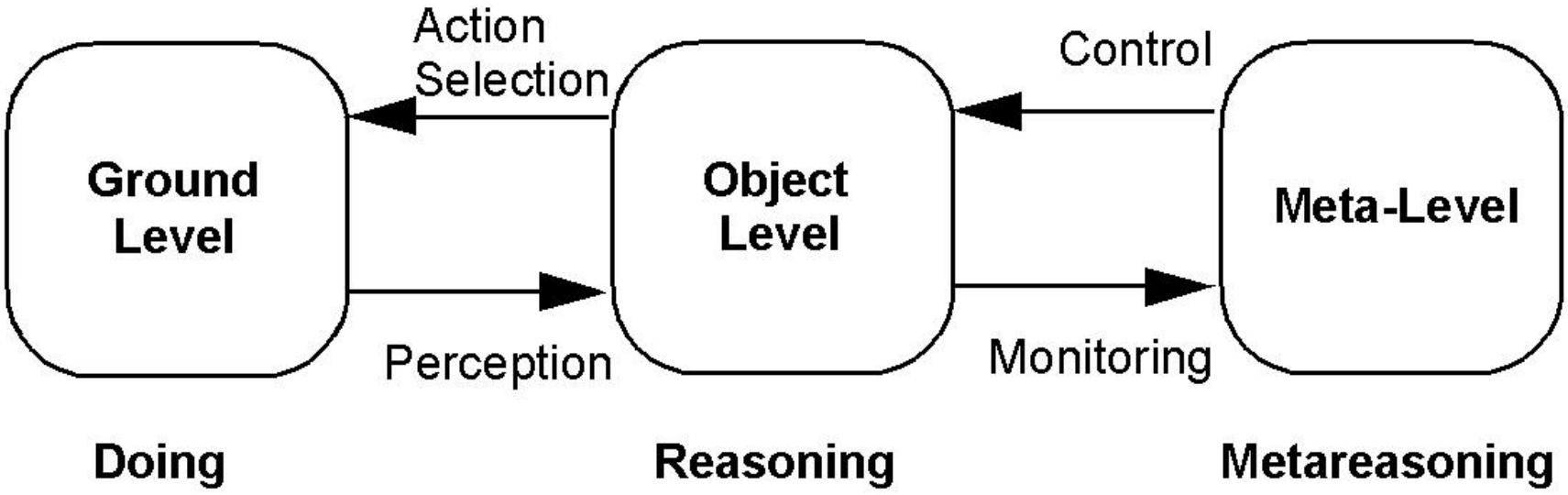
BBN TECHNOLOGIES

Metareasoning, Monitoring and Self-Explanation

Michael T. Cox
mcox@bbn.com http://mcox.org
Intelligent Computing
BBNT, Cambridge

Duality in Reasoning & Acting



Meta-AQUA System

- Performance Task: Story Understanding
  - Explanation of anomalous events
  - No Agent Action or Meta-Level Control
- Learning Task:
  - Input trace of reasoning leading to failed explanation
  - Reason about trace to explain failure
  - Generate learning goal to fix knowledge
- Meta-Explanation:
  Explaining Explanation Failure

Meta-AQUA Example

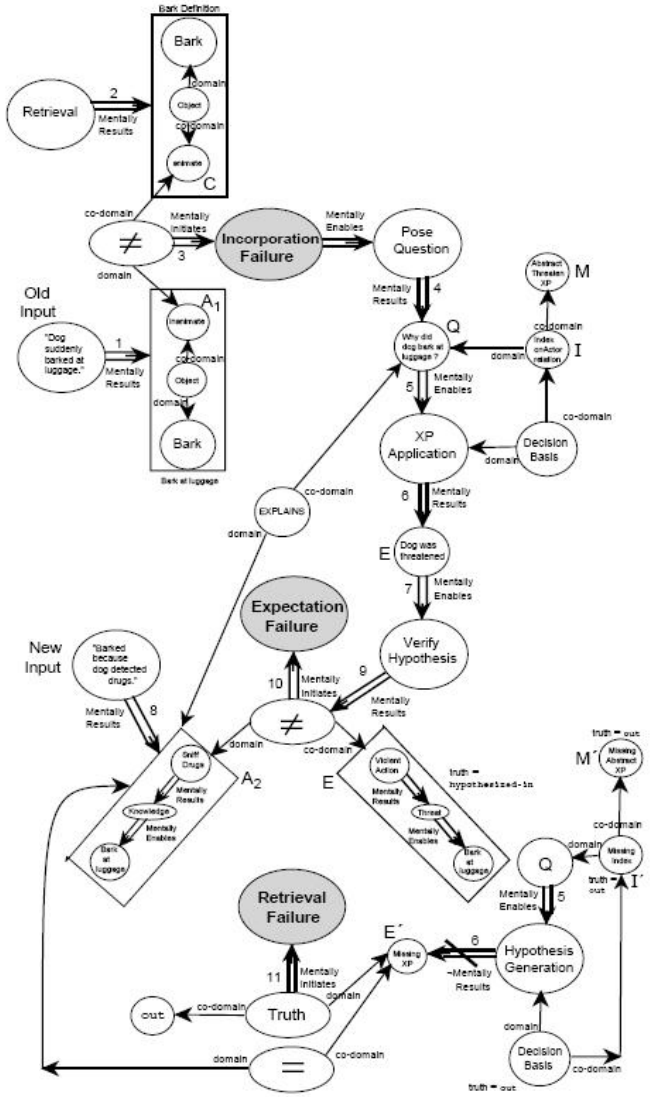
S1: A police dog sniffed at a passenger's luggage in the airport terminal.

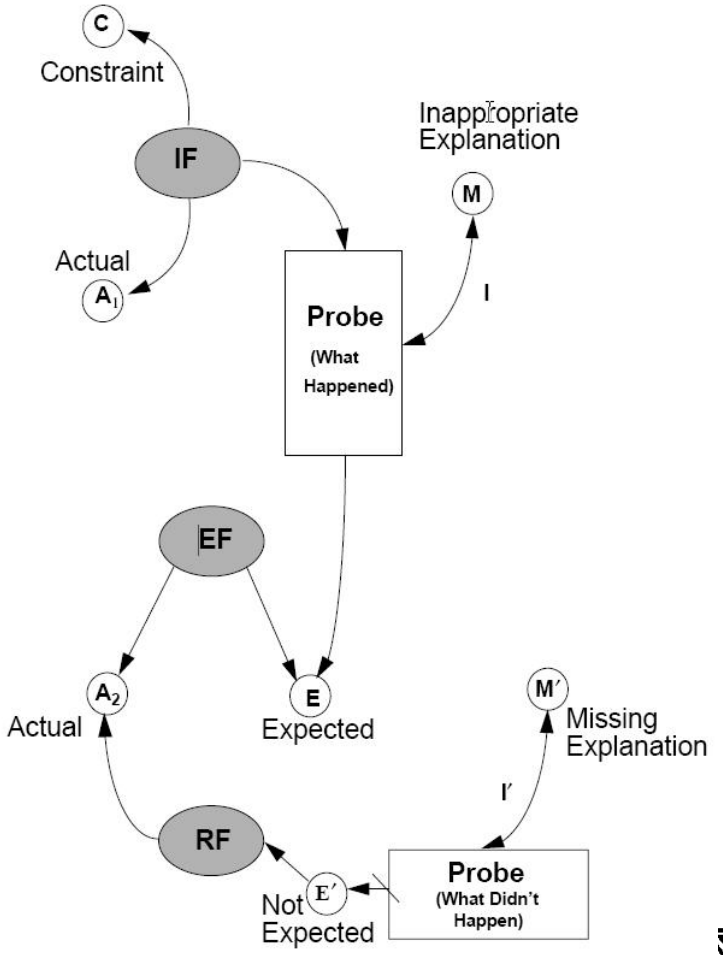
S2: The dog suddenly began to bark at the luggage.

S3: The authorities arrested the passenger, charging him with smuggling drugs.

S4: The dog barked because it detected two kilograms of marijuana in the luggage.

Meta-AQUA Demonstration

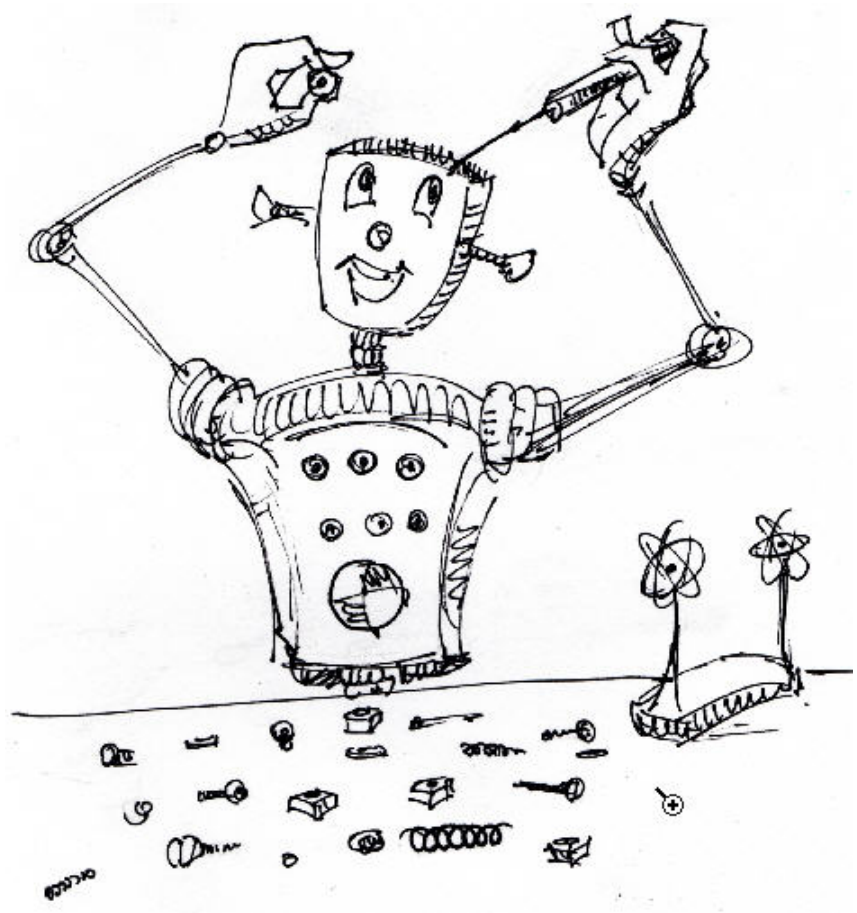




4 Themes for Meta-Reasoning in Agent-Based Systems

1. Self-Modifying Code
2. Self-Knowledge
3. Self-Understanding
4. Self-Explanation

Self-Modifying Code



- Direct Control of performance to change the code
- Indirect Control to change the knowledge
- But to change the agent function is what?

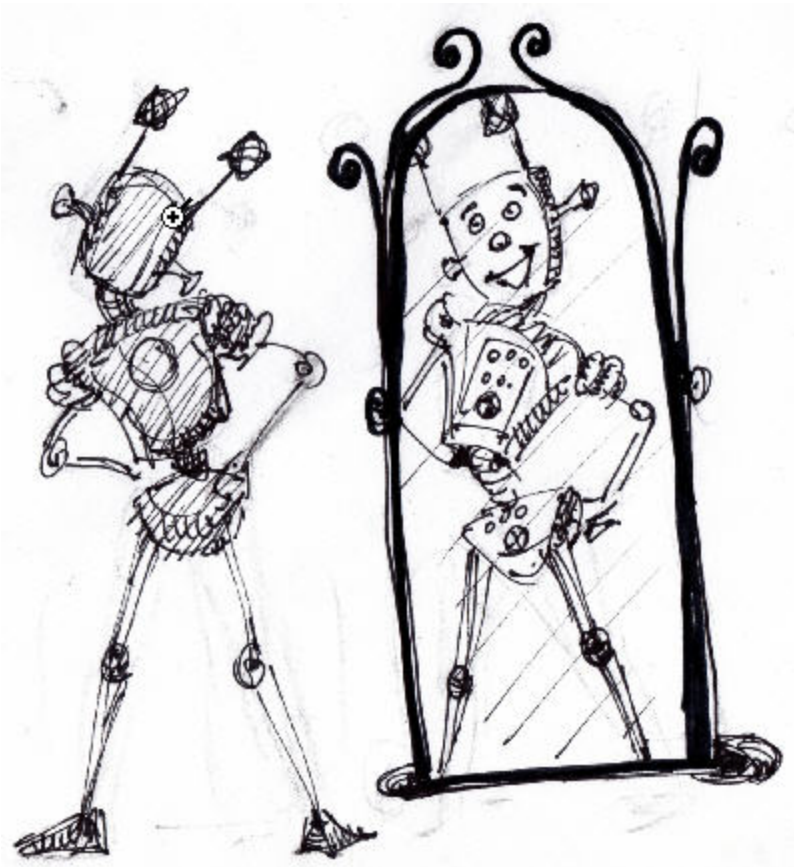
Self-Modifying Code

"Once self-description is a reality, the next logical step is self-modification. Small, self-modifying, automatic programming systems have existed for a decade; some large programs that modify themselves in very small ways also exist; and the first large fully self-describing and self-modifying programs are being built just now. The capability of machines has finally exceeded human cognitive capabilities in this dimension; it is now worth supplying and using meta-knowledge in large expert systems."

— Lenat, Davis, Doyle, Genesereth, Goldstein and Schrobe, 1983 (p. 238)

Self-Knowledge

- Knowledge about the Self
  - Confidence factors
  - Episodic memory
- A difference exists between
  1. Metaknowledge
  2. Metareasoning



Knowledge and Process

Metaknowledge

- Knowledge about knowledge

- Planning is cognitive
- Plans are knowledge produced by mental process

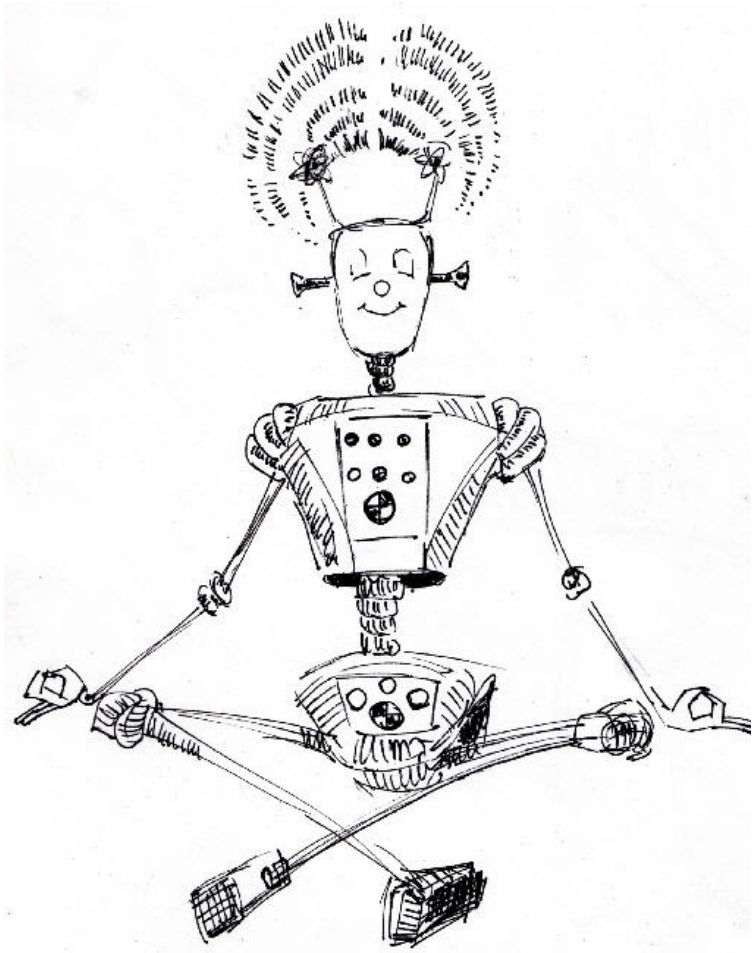
- "P is a good plan."
  - Assertion about P
  - Metaknowledge

Metareasoning

- Reasoning about Reasoning

- Planning is cognitive
- Process can be represented in trace
- Reasoning about the trace is metacognitive
- "The planning was poor."
  - Assertion about process that produced P
  - Metareasoning

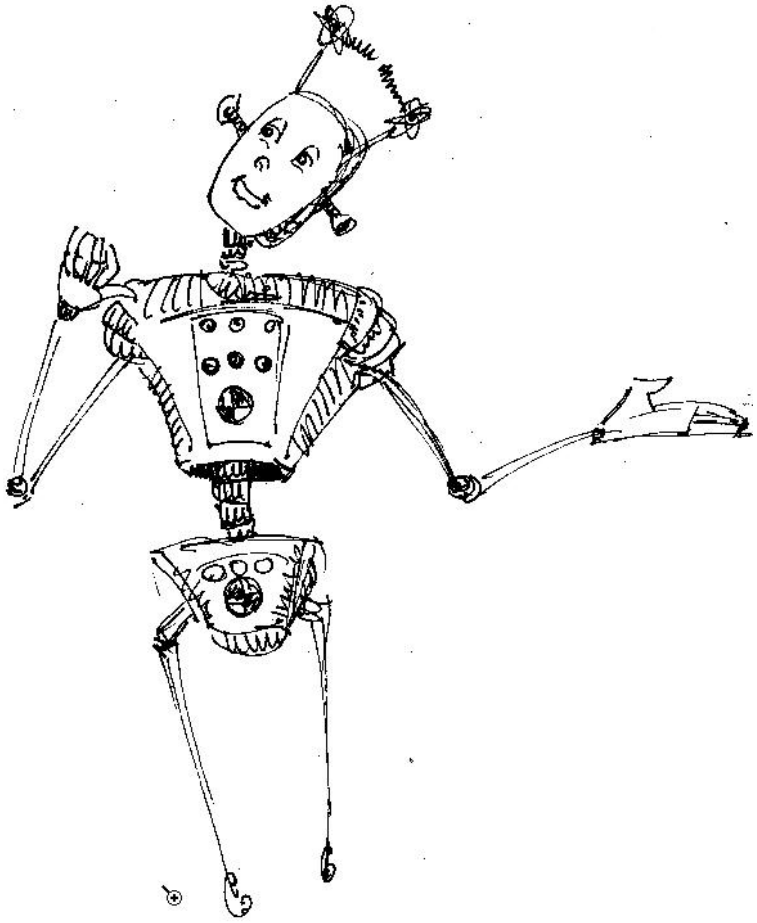
Self-Understanding



- What does it mean to understand a story?
  - Find and adapt a schema in memory that represents a coherent causal interpretation of events and change.
- What does it mean to understand the self?
  - Self-interpretation

Self-Explanation

- **Explanation**
  provides answers to *how* and *why* questions in the world.
- **Self-Explanation**
  - Maps *symptoms* of cognitive failure to failure *causes* in the head.
  - Points to necessary knowledge changes and learning.



10 Basic Mental Explanations

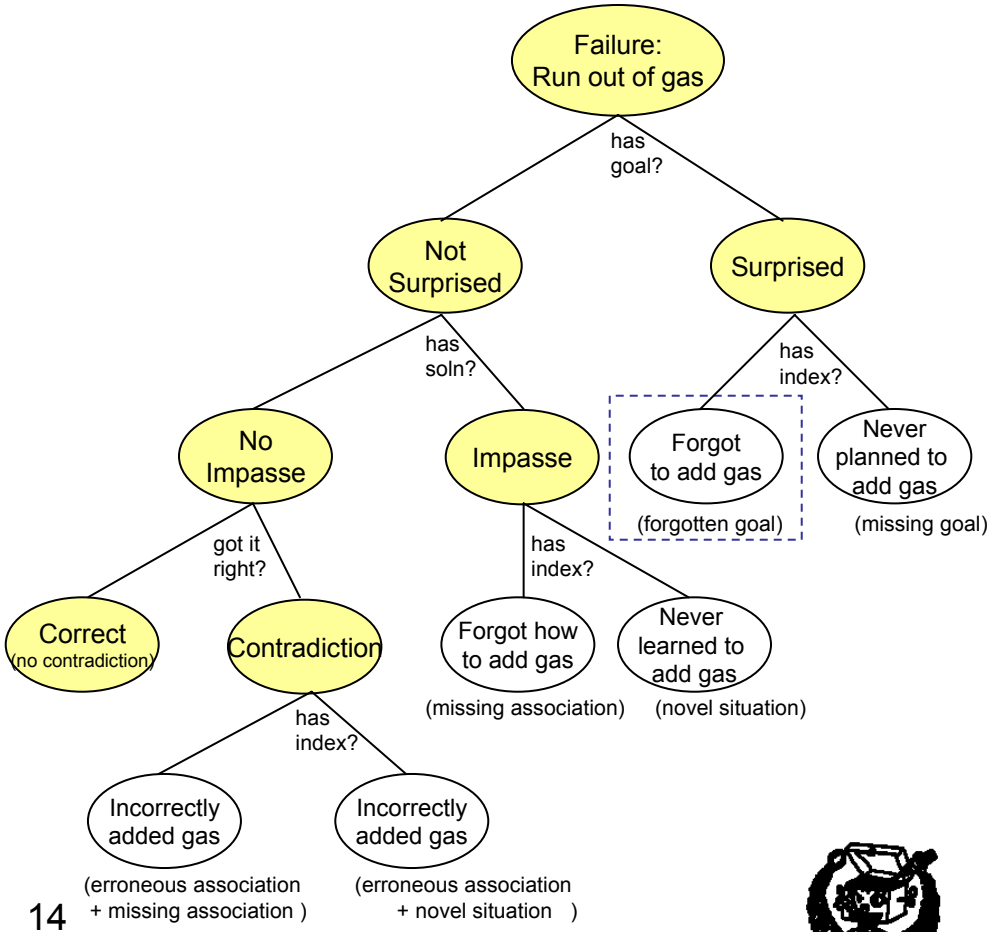
1. I forgot that X.
2. I am good at Y.
3. I did not see (or notice) Z.
4. I mistook an M for an N.
5. I assumed that I is the case because B.
6. I thought that all J could K.
7. I learned that Q today.
8. I did not have enough time to think about R.
   I wasted time worrying (thinking) about R.
9. S surprised me because B.
10. I chose to do A1 instead of A2 because B.
    I wanted to achieve G1 rather than G2 because B.

How Self-Explanation Helps the System Learn

- **The performance task**: Navigation plan execution.
- **The failure symptom:** *Surprised* when vehicle runs out of gas.
- **The failure cause:** *Forgot** to fill gas tank.
- **The learning:** *Remember* to check the gauge before driving.

* But will cognitive systems forget? See Cox, M. T. (1994). Machines that forget: Learning from retrieval failure of mis-indexed explanations. In *Proceedings CogSci-94* (pp. 225-230).

Space of cognitive failure

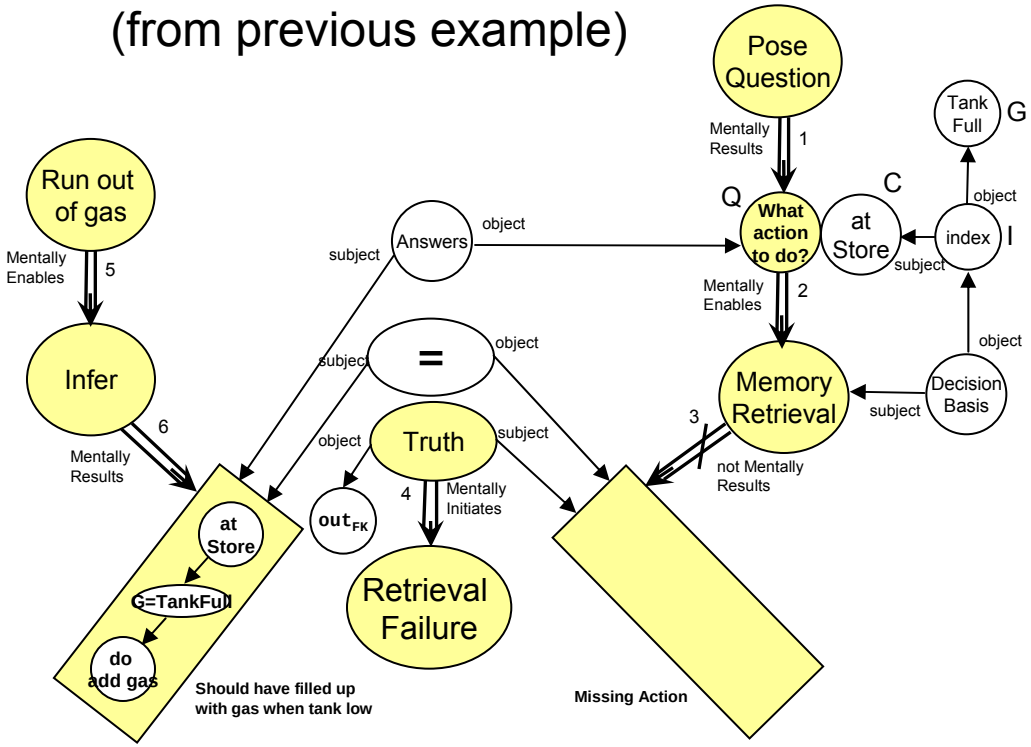


BBN TECHNOLOGIES

How Self-Explanation Helps the User Understand the System

Graph Input

(from previous example)



"**Forgetting to fill up with gas**"
Meta-XP eXplanation Pattern Structure.

Text Output

Paraphrase: "I forgot to fill up the car with gas when I was at the store."

Elaboration: "The context, C, of being at the store did not sufficiently match the index, I, with which the goal, G, to have a full tank was stored in memory, so I failed to retrieve the goal at the right time and thus did not put gas in the tank. Because the tank was low, I did not have enough fuel and then ran out of gas."

Conclusion

- Challenge to the Agent Community interested in Metareasoning
  - Monitoring as well as control is a first class citizen
  - Self-explanation can be a part of the evaluation
- Learning as an explanatory debugging dialogue
  - Statistics alone not enough
  - Can understand what went wrong during plan execution by relating mental factors such as memory, emotion, knowledge, and inference to flawed action choices.
- Integrated cognition:
  - Problem solving + interpretation + learning.
- Integrated metacognition:
  - Control of cognition + monitoring of cognition.
  - Metaknowledge + self-knowledge