OMSCS 7637/ CS 7637/ CS 4635

Ashok Goel

Fall 2016

Final Examination

Assigned Wednesday, November 30, 2016

Due Thursday, December 8. 2016

This is a take-home examination. You may consult any book, video, website, etc. However, your answers must be your own work: you may not consult any person.

A good length for an answer is about 2000 words. However, this is neither a minimum requirement nor an upper limit. We are more interested in the quality of the answer than in its length.

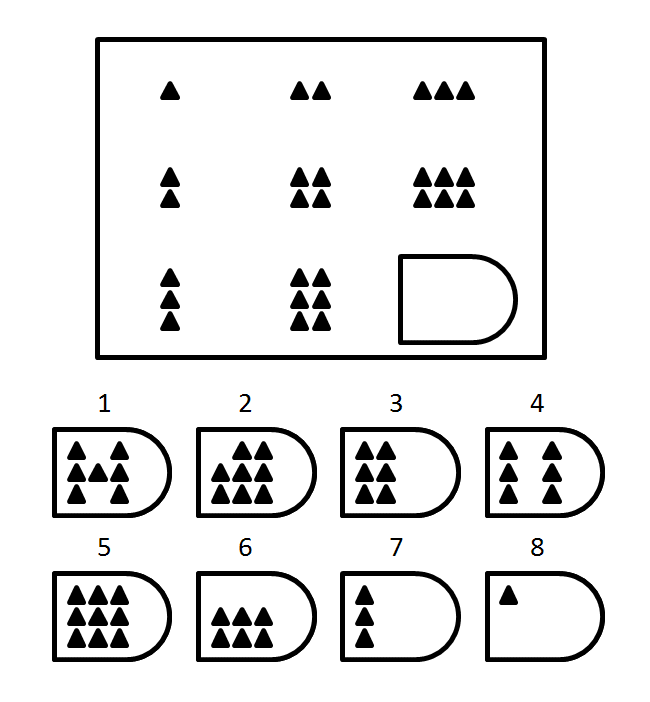
Trust is becoming an increasingly major issue in AI. Can we trust AI systems? On what basis might we trust an AI system? If humans cannot trust AI systems, then will humans ever accept AI agents in everyday work, play, life?

Explanation is said to be one way of engendering trust. If AI systems could only explain themselves, the argument goes, humans are more likely to trust them. That is, if an AI agent could not only give a result, but also explain how its knowledge and reasoning led it to the result, then humans are more likely to accept the result.

Unfortunately, many AI methods are opaque and do not foster trust. This includes many machine learning techniques for AI (MLAI) such as neural networks and deep learning. Thus, there are increasing efforts to add the capacity for explanation to such methods. (See, for example, <http://www.darpa.mil/program/explainable-artificial-intelligence>).

Fortunately, knowledge-based techniques for AI (KBAI) are very suitable for generating explanations in part because they represent knowledge explicitly. Consider for example the AI agents you yourself constructed for Projects 1 and 2 in our KBAI class. Your agent for Project 1, for example, likely explicitly represented its knowledge of geometric objects such as triangles and circles, explicitly represented the configuration of objects in an image as a semantic network, explicitly calculated the similarity between two images, explicitly selected an answer based on the similarity calculations, etc., etc. Thus, your agent should be able to explain its reasoning and how it leads to the selected answer.

This seems a little similar to human behavior in generating explanations. Consider, for example, the problem in Figure 1 with #5 as the correct answer:



**Figure 1: An Illustrative Problem**

I could ask you several questions about your own problem solving such as:

1. What is the object in the eight images displayed in the matrix?
2. What is the relationship between the three images in the top row of the matrix?
3. What is the relationship between the three images in the middle column?
4. What is the difference between the #3 and #4 as potential answers?
5. Why did you select #5 as the answer?
6. Why is #3 not the correct answer?

I am sure you yourself can easily generate good answers – explanations – for each of these questions. We want the agents you constructed in the KBAI class to explain themselves as well.

One way of doing this is to add a layer of metacognition to your agents. At present your agents are capable only of deliberation: memory, reasoning and learning. But they don't *know* anything about their ownmemory (or knowledge), reasoning and learning, and thus they cannot explain themselves. But let us suppose you could add a metacognitive layer to your agents such that the metacognitive layer could inspect the memory (knowledge), reasoning and learning in the deliberative layer to generate explanations.

Figure 2 illustrates this architecture (please ignore the reactive layer in the figure for this examination).

Cognitive System

Reaction

Deliberation

Metacognition

Input

Output

**Figure 2: Agent Design**

Your task on this final examination is to provide a detailed design of a KBAI agent that can generate explanations. The agent should have the high-level design shown in Figure 2 (again ignore the reactive layer). It should be capable of answering the six questions listed above for the problem illustrated in Figure 1.

Here are some hints that may help you: First, write down the answers that you yourself may generate for the six questions listed above. Second, write down the knowledge, reasoning and learning your agent uses in the deliberative layer (like you did in the reflection reports). Third, add an explanation interface to your agent that may allow a human to pose questions to your agent. Finally, show how meta-cognitive introspection on the deliberative reasoning helps generate the right answers to the input questions. In particular, show how your agent will understand the question, how your agent will introspect on the memory (knowledge), reasoning, and learning in the deliberative layer, and how it will then generate answers to the input questions.