Verification

Rich Sutton

11/14/2001

If the human designers of an AI are not to be burdened with ensuring that what their AI knows is correct, then the AI will have to ensure it itself. It will have to be able to verify the knowledge that it has gained or been given.

Giving an AI the ability to verify its knowledge is no small thing. It is in fact a very big thing, not easy to do. Often a bit of knowledge can be written very compactly, whereas its verification is very complex. It is easy to say "there is a book on the table", but very complex to express even a small part of its verification, such as the visual and tactile senations involved in picking up the book. It is easy to define an operator such as "I can get to the lunchroom by going down one floor", but to verify this one must refer to executable routines for finding and descending the stairs, recognizing the lunchroom, etc. These routines involve enormously greater detail and closed-loop contingences, such as opening doors, the possibility of a stairway being closed, or meeting someone on the way, than does the knowledge itself. One can often suppress all this detail when using the knowledge, e.g., in planning, but to verify the knowledge requires its specification at the low level. There is no comparison between the ease of adding unverified knowledge and the complexity of including a means for its autonomous verification.

Note that although all the details of execution are needed for verification, the execution details are not themselves the verification. There is a procedure for getting to the lunchroom, but separate from this would be the verifier for determining if it has succeeded. It is perfectly possible for the procedure to be fully grounded in action and sensation, while completely leaving out the verifier and thus the possibility of autonomous knowledge maintainence. At the risk of being too broad-brush about it, this is what typically happens in modern AI robotics systems. They have extensive grounded knowledge, but still no way of verifying almost any of it. They use visual routines to recognize doors and hallways, and they make decisions based on these conclusions, but they cannot themselves correct their errors. If something is recognized as a "doorway" yet cannot be passed through, this failure will not be recognized and not used to correct future doorway recognitions, unless it is done by people.

On the other hand, once one has grounding, the further step to include verification is less daunting. One need only attach to the execution procedures appropriate tests and termination conditions that measure in some sense the veracity of the original statement, while at the same time specifying what it really means in detail. What is a chair? Not just something that lights up your visual chair detector! That would be grounded knowledge, but not verifiable; it would rely on people to say which were and were not chairs. But suppose you have routines for trying to sit. Then all you need for a verifier is to be able to measure your success at sitting. You can then verify, improve, and maintain your "sittable thing" recognizer on your own.

There is a great contrast between the AI that I am proposing and what might be considered classical "database AI". There are large AI efforts to codify vast amounts of knowledge in databases or "ontologies", of which Doug Lenat's CYC is only the most widely known. In these efforts, the idea of people maintaining the knowledge is embraced. Special knowledge representation methods and tools are emphasized to make it easier for people to understand and access the knowledge, and to try to keep it right. These systems tend to emphasize static, world knowledge like "Springfield is the capital of Illinois", "a canary is a kind of bird", or even "you have a meeting scheduled with John at 3:30", rather than the dynamic knowledge needed say by a robot to interact in real time with its environment. A major problem is getting people to use the same categories and terms when they enter knowledge and, more importantly, to mean the same things by them. There is a search for an ultimate "ontology", or codification of all objects and their possible relationships, so that clear statements can be made about them. But so far this has not proven possible; there always seem to be far more cases that don't fit than do. People are good about being fluid with there concepts, and knowing when they don't apply.

Whatever the ultimate success of the symbolic "database AI" approach, it should be clear that it is the anti-thesis of what I am calling for. The database approach calls for heroic efforts organizing and entering an objective, public, and disembodied knowledge base. I am calling for an AI that maintains its own representations, perhaps different from those of others, while interacting in real time with a dynamic environment. Most important of all, the database approach embraces human maintainence and human organization of the AI's knowledge. I am calling for automating these functions, for the AI being able to understand its knowledge well enough to verify it itself.