SWCP Portal
February 2016

When androids dream of electric sheep

Can Robots Have Souls?

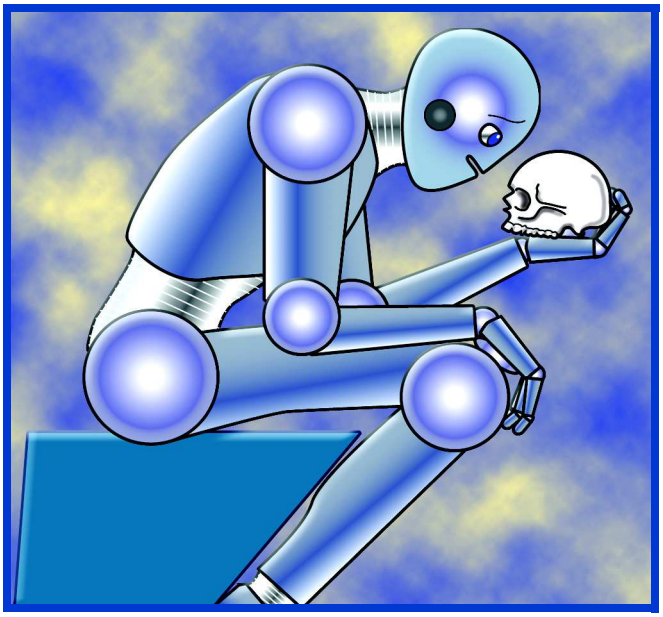
There are exciting developments in the futuristic field of **Artificial Intelligence** (AI) and **robotics**. Decades of dead ends and detours made it seem as if smart, self-aware machines might forever remain the stuff of science fiction. Quite a few came to doubt that consciousness could ever be put into a can.

But recent surprising breakthroughs herald success. AI is already being put to work in the real world, in smartphones, search engines, driverless cars, and that's just the beginning. It is raising many questions.

From ancient legends to modern labs

The dream of mechanical replacements for human labor is very old. But robots are more than windmills or other tools; they must be smart enough to react to their environment. First imagined as metal men, other forms followed surprisingly early on.

Robots appeared in Greek mythology with **Talos**, the giant bronze guardian of Crete. The lame god **Hephaestus** (called Vulcan by the Romans) created golden women shop assistants much like C3PO and also table-topped tripods used to run messages up to Olympus. The direct ancestors of R2D2, these were non-**humanoid robots** and in essence, networked. Only now are these finally being realized.



The first real humanoid **automaton**, however, may be the armored knight built by Leonardo da Vinci in 1495. His investigations, and those of other Renaissance anatomists, soon uncovered how muscles leverage bones and the workings of bodily structure.

Throughout the age of clockwork automatons, mechanics steadily improved. Victorians developed **feedback devices** to regulate steam engines as well as electric **servomotors** for precise motion control. Such advances eventually led up to **ASIMO**, a robot that can walk and even dance, debuting in 2000.

Sensor schemes modeling vision, hearing, and touch also had to be devised, of course, and are still being developed. Not until the advent of digital computers last century did brains finally catch up with bodies.

The real trick, how to mimic thought, long remained elusive. AI's first successes came with **expert systems** where specialized knowledge fed into data banks was applied to problems. This let computers play chess but didn't make them winners, as foes came up with new moves that were not in their playbooks.

And simulating real situations like ordering in a restaurant, while following patterns easily understood by people, proved that scripting all possible responses to novel events was utterly impractical.

So the scientists decided to throw massive computing power at the problem. Perhaps just a big enough brain was needed. Predicting any potential outcomes of moves enabled computers to up their skills and IBM's **Deep Blue** finally beat a grandmaster in 1997.

But even with the brute force of supercomputers, the real world was too complicated. To make sense of alternatives, programmers turned to **big data** – vast arrays organized on servers composed of **GPUs** – chips designed to control game graphics, which proved perfectly suited to the task. They used this data to allow the machines to teach themselves.

So, to instruct computers how to recognize cats from dogs, show them thousands of pictures, and allow them to work it out on their own with minimal direction. This is much the way humans learn, but just as children need guidance, so too do the machines.

A picture may be worth a thousand words but mistakes of interpretation will happen. The computer, for

Continued on back

example might come to the conclusion that a collar is a vital part of the kitty and not recognize it without one. Error-correction techniquses must be employed.

This might be a critical limit for this method. The more complex a concept, the more likely a mistake will occur, and the more important it is, the worse the consequences of errors will be. Computers can be trained to tell cats from dogs, but could they ever understand "love," even with a million pictures?

Gaming the system

Even so, this so-called "**deep learning**" strategy has scored impressive results. Lately a computer built by Google beat a grandmaster at **Go**, an Oriental board game, which although simpler than chess, is much more subtle and often **considered** something machines might not achieve for decades more.

Google's **DeepMind** team started by first programming their Go-bot with a library of 30 million moves from expert Go players. Then they played it *against itself,* and by doing it again and again with new moves and slightly different assumptions, were finally able to train their AI to truly master the game.

The role of games in all this is *not* coincidence. Gaming provides limited arenas with clearly-defined rules and variability governed by strict probability. Moreover, it allows machines ways to deal with uncertainty. After all, guessing is a vital strategy for how we survive in our goosey-loosey world, but it's something that digital devices are poorly suited to do.

Monte Carlo localization, for instance, allows a robot to judge its location by generating an array of possible positions. It constantly re-evaluates them from sensor readings as it moves by means of **Bayes' law** which is a means of estimating how likely it is to really be in that spot from other probabilities.

Again, this is rather like what we do unconsciously all the time. But it may be disconcerting to realize that robots and AI systems are not any more absolutely accurate than we are – and in some ways far *less* so.

The face in the mirror

Can computing systems, no matter how smart, be anything more than **glorified toasters**? **Alan Turing**, "the father of AI", long ago proposed that a machine that could fool a human being in a conversation was truly intelligent. While some early systems, like **ELIZA**, were able to engage humans into revealing intimate details of themselves by repeating statements as questions, **not even** the iPhone's Siri personal assistant has yet beaten the **Turing test**.

But robots have already demonstrated sentience. One experiment was based on an ancient puzzle called the "**King's Wise Men**". In this **demonstration**, three cute little **NAO robots** are told in effect that two were given a "dumbing pill" preventing them from speaking, but that one got a fake. When asked which pill they took, one droid slowly gets up and answers "I don't know". It then *realizes* that it spoke and that this means that it didn't get the dumbing pill. So the little bot politely corrects itself, raising its hand and replying that it *does* know, now.

Wow. That's at least *three* stages of cognition: understanding its inner state, relating it to the world, and correcting its self-evaluation. Sure, the realistic gestures help but the result looks like real self-awareness rather than a clever simulation. Our consciousness of ourselves is an integral part of the human condition; perhaps this demonstrates that computers are even now naturally evolving in the same direction. And then what? Is it right to ever pull the plug?

However, self-awareness is *not* all there is to a person. Metaphysical elements aside, emotions, attitudes, idiosyncrasies, and morality all go into molding a personality. And all those, as actors, politicians, and con artists continually prove, *can* be simulated.

The machines are quickly learning how. **Social robots**, those designed to interact with us (C3PO again), will specialize in understanding the nuances of human communication and responding in an appropriate manner. But that may not always be sweet and demurely obedient like good servants. Already they're being taught how to simulate **anger** and even how to **say "no"** for their own safety.

In time, something like Asimov's famous **Three Laws of Robotics** might be instituted to keep our metal slaves docile. Without living flesh to govern them through pain, pleasure, and feelings, robots must remain radically different from us. But they might get some kind of cold satisfaction from their programming in fulfilling orders or cleaning up our messes.

Who cares as long as they don't cause trouble? But creating self-aware entities carries responsibilities. Like the Internet, robots are a mirror of our soul. What we focus on will determine the outcome. Whether they'll be servants, companions, or masters, as their "parents" we *owe* it to the bots to raise them properly.