OkCupid and Your Mechanical Friend

February 22, 2011 AI, images
0

By: Joshua Fox

When the first artificial general intelligence is built, we want it to be our friend.

If it is not, we are in big trouble. That's because an AGI will quickly become ultraintelligent: Whatever its goals, improving its own intelligence is a good a way of achieving them, and it won't stop until it is much smarter than humans.

An ultraintelligent robot which hates us is a bad thing: We can't possibly outsmart it.

An ultraintelligent robot which doesn't care about us is also a bad thing: It will probably wipe us out by consuming all available resources, not out of malice, but simply because it ignores our well-being as it works towards its goal. Though we don't know much about the methods it may use, we do know one thing: It will almost certainly achieve its goal.

An ultraintelligent robot which loves us is a very good thing.

There's no reason that an intelligent machine needs to hate us, or love us. Hate and love are complex and very human emotions, brewed over eons in our environment of evolutionary adaption. Hate includes envy, resentment, a lust for power, enmity for other tribes; love includes sublimation of the self, altruism, affection for family members, and lust of a better kind.

These emotions are complex, and messy. Humans are contradictory animals. Love for one's tribe is reinforced by hate for another tribe. The two lusts all-too-often coexist. Love-hate relationships remind us of the messy swirl of motivations which makes us who we are.

A future AGI might be copied from a human model, or it might be based partially on the human model, including emotions like love and hate. But it doesn't have to be. An AGI is defined as a type of optimizer, an entity which seeks to bring the world towards some goal. If the AGI achieves complex goals in complex environments, we say that it is a powerful optimizer. The goal could be anything. In Bostrom's thought-experiment, the goal is to accumulate as many paper-clips as possible. Or, to take some more likely examples, the goal could be winning at chess, winning at war, making money, curing cancer, or making the human race better off forever.

And to achieve these goals, the AGI doesn't need to be structured anything like a human brain. I don't know how the AGI will be structured, but I usually imagine a data center, with computer chips churning away; thousands of lines of soulless software code executing cold, hard algorithms. No consciousness, no self, no feelings. A look at the C source code of an AGI software project like OpenCog, for example, is a good reminder that an AGI need not be anything like humans. (Not that OpenCog will necessarily be the first AGI to take off to ultraintelligence.)

Computers already work for us today. Sometimes they have bugs, but they do more or less what we want. But more or less isn't good enough for an ultraintelligent computer. It needs to be focused on exactly the right goal, because the goal is exactly what's going to happen. We have no chance of outwitting the AGI or locking it up if getting what what we think we wished for turns out to be an unpleasant surprise. The AGI's goal must be to give humans what they really and truly want; not just food, or sex, or smiles, but the most spiritual, most exalted desires of the human race. Coherent Extrapolated Volition is one proposal for this. Coherent Aggregated Volition is another.

But is that possible? Can an unhuman computer really optimize for the deepest human values? Computers today work towards basic human values like getting food on the table, or defeating the enemy: They direct

the food supply chain and target weaponry. But can AGIs really optimize towards the most sublime, most meaningful, human values like love, creativity, freedom, exploration, growth, and happiness?

Sure they can. It already happens today.

Look at OkCupid, the dating site which built its reputation with a statistics-aware blog. OkCupid's software has the goal of increasing love in the world. It's not AGI, but it has some fancy narrow AI, and makes a good example of a machine which optimizes for one of humanity's deepest desires.

OkCupid's not perfect, but I think that on balance it adds love to the world, rather than decreasing it.

It's nothing but a heartless automaton spinning away on the CPU's, and yet it helps us achieve towards one of humanity's most important values, just a little.

A short digression on the topic of goals: An intelligence works to its goals, but these need not be the true goals of its creators. The real metric for the OkCupid software is not actually love; it works hard to maximize the number of couples which get together in relationships or marriage; the software can't measure true love. This works fine for a narrow AI, but an ultraintelligent OkCupid with that metric would tweak humanity's brains to pair off on the spot, love be damned. This is a problem with the goal, not with the optimization power of the intelligence. When, in coming decades, an AGI is ultimately built, we need to specify its goals to be exactly what we want, not an approximation.

Pairing people off is the algorithm's direct goal. But the high-level goal for which the software was built is to make the founders some money, and in fact, they sold out for a nice bundle. But the software need not know that. Like any optimizer, it works towards the end-goal which is set for it, not a goal which it figures out for itself

OkCupid is already superhuman in some ways. It can match up a lot more couples than any human matchmaker, and for a lot less money–the denomination of our resources–than any number of human matchmakers. Still, it's not an AGI: It does not have a flexible, general intelligence. We can hope for a future OkCupid with the smarts of a human and beyond, one which would do a far better job than the machines of today. On the other hand, might it be that heartless software will hit a hard ceiling in the realm of love? Maybe only an entity with emotions, with feelings, with the complex mix of quirks which make us human, could truly excel as a Yenta?

Yet an optimizer doesn't need to be what it optimizes. Software which optimizes petroleum production schedules is nothing like the systems which actually run the oil company. Optimizers process variables which have been carefully extracted from the systems being optimized.

On the other hand, some optimizers do simulate: They run a copy of the system being optimized. The copy can be simplified, or a full emulation. Some software today does this, and even humans mirror the thoughts of other humans in their minds to try to figure out what they want. An ultraintelligent AGI could simulate humans. But even this machine doesn't need to be human. The simulations can run in a tiny part of its mind, toy worlds for it to observe, while it looks on from above, the god of its own internal domains, charting the best path to its goals using whatever incomprehensible and perhaps very unhuman mental architecture it may have.

Remember, OkCupid is bringing love to humans, not to itself. A digger digs holes for humans, it doesn't need a basement. An travel search engine finds the cheapest flight for you; it's staying put. Don't be ego-centric: Just because you have some personal goals–life, love, creativity, exploration, whatever–doesn't mean that the AGI has the same ones for itself. The AGI has whatever goal it's been designed with, or which it stumbled into as the result of bugs. If the goal is to help humans, all the better for us.

To paraphrase Eliezer Yudkowsky in an H+ Magazine article: It doesn't love you; it doesn't hate you; it's just doing a great job of making your life much, much better.

Your mechanical friend can be your friend and stay mechanical through and through, and that's a good thing. Do you really want your ultraintelligent servant to get personally involved? Yenta in Fiddler on the Roof was enough trouble, even when trying to help; now, imagine her a million times smarter than you. Humans are

complex and contradictory creatures, full of biases and conflicting goals. Sometimes they're on your side, sometimes not.

Wouldn't you rather have a helper full of wits, wisdom, cleverness, and guile, but without any feelings or goals of its own, other than to give you what you truly, deeply, want?

Joshua Fox works at IBM, where he co-founded the Optim Data Redaction product and now manages its development. He has served as a software architect in various Israeli start-ups and growth companies. Fox speaks and writes for business, technical, and academic conferences and journals. He received his PhD from Harvard and his BA summa cum laude from Brandeis. Links to his talks and articles are available online. (Image credit: http://www.flickr.com/photos/skreuzer/354316053/)