Printed: 2016-05-06

**Institute for Ethics and Emerging Technologies**

---

IEET Link: http://ieet.org/index.php/IEET/more/loosemore20121128

**The Fallacy of Dumb Superintelligence**

**Richard Loosemore**

***Ethical Technology***

http://ieet.org/index.php/IEET/ieetblog

November 28, 2012

This is what a New Yorker article has to say on the subject of "Moral Machines": "An all-powerful computer that was programmed to maximize human pleasure, for example, might consign us all to an intravenous dopamine drip."

What they are trying to say is that a future superintelligent machine might have good intentions, because it would want to make people happy, but through some perverted twist of logic it might decide that the best way to do this would be to force (not allow, notice, but force!) all humans to get their brains connected to a dopamine drip.

I have been fighting this persistent but logically bankrupt meme since my first encounter with it in the transhumanist community back in 2005. But in spite of all my efforts .... there it is again. Apparently still not dead.

Here is why the meme deserves to be called "logically bankrupt".

If a computer were designed in such a way that:

(a) It had the motivation "maximize human pleasure", but

(b) It thought that this phrase could conceivably mean something as simplistic as "put all humans on an intravenous dopamine drip", then

(c) This computer would NOT be capable of developing into a creature that was "all-powerful".

The two features <all-powerful superintelligence> and <cannot handle subtle concepts like

"human pleasure"> are radically incompatible.

With that kind of reasoning going on inside it, the AI would never make it up to the level of intelligence at which the average human would find it threatening. If the poor machine could not understand the difference between "maximize human pleasure" and "put all humans on an intravenous dopamine drip" then it would also not understand most of the other subtle aspects of the universe, including but not limited to facts/questions like:

"If I put a million amps of current through my logic circuits, I will fry myself to a crisp",

or

"Which end of this Kill-O-Zap Definit-Destruct Megablaster is the end that I'm supposed to point at the other guy?".

Dumb AIs, in other words, are not an existential threat.

This myth about the motivational behavior of superintelligent machines has been propagated by the Singularity Institute (formerly known as the Singularity Institute for Artificial Intelligence) for years, and every one of my attempts to put a stop to it has met with scornful resistance. After so much time, and so much needless fearmongering, I can only conclude that the myth is still being perpetuated because it is in the Singularity Institute's best interest to scare people into giving them money.

We should certainly talk about the threats posed by smart technology, but at the same time we need to make sure we really are talking about smart technology, not some hypothetical Dumb-AI system that are is clueless that it could never make it up to the level of human level intelligence, never mind the level of humanity-threatening superintelligence.

Objections

Let's take a look at some of the objections to the line of argument I have just presented.

You might ask "Aren't powerful dumb things at least as threatening as powerful smart things?"

That seems reasonable at first sight, but it turns out to be a loaded question. Yes, it is true that powerful dumb things are as threatening as powerful smart things -- but my argument is that when a computer is as dumb as that, it cannot get to be as powerful as that. If the AI is (and always has been, during its development) so confused about the world that it interprets the "maximize human pleasure" motivation in such a twisted, logically inconsistent way, it would never have become powerful in the first place. Powerful dumb things never get to be superintelligent powerful dumb things.

But now, what about the kind of military AI already in use, in drones and other weaponized hardware? The answer is that nothing in this argument is meant to address the very

different problem of "quite powerful" machines that do dumb things. We already have those, and more of them are being built by the minute. Drones that are set to kill with no human in the loop, for example, are dangerous, but they are not existential-threat dangerous. A drone might get out of control and cause immense carnage, but if the worst comes to the worst you can just wait for it to run out of fuel.

Today's dumb AI technology does need to be discussed, but I have to leave that question for another day because the myth I am attacking is specifically the one that refers to "all-powerful" AIs (as the New Yorker article phrased it), and the kind of motivations that they might have.

Another objection is that in spite of my protestations it really might be possible for a system to be so intelligent that it could outsmart all of humanity, while at the same time being so uninitelligent about matters of motivation and goals that it could think that a dopamine drip should be forced on humans to make them happy. This is a fair point, but the trouble is that such a claim has so many prima facie problems that it begs a huge number of questions. To date, I have seen nothing in the way of a detailed explanation for how such a contradictory situation could arise. Instead, I see just a blanket assumption that it is "obviously" possible, or "obviously" inevitable. So obvious that many people feel there is no real need for supportive reasoning. My attack, then, is directed at this automatic assumption that such a scenario is a reasonable possibility, in spite of its obvious, massive internal inconsistency.

Higher Form of Reasoning?

Here is another possible objection. "Isn't part of the idea that at some 'higher level' of reasoning a dopamine drip might make more sense than our chaotic thinking about happiness?"

Well, that seems like a pretty strange "higher form of reasoning", but let's entertain it as a possibility, and see if we can understand what that kind of reasoning must entail.

First, though, we have to be absolutely clear about what the premise is. For the AI to come to the conclusion that "maximize human pleasure" means that it must "consign us all to an intravenous dopamine drip", the AI would have to be so narrow-minded as to think that maximizing human pleasure is a single-variable operation (thereby rejecting a vast swathe of human thought pertaining to the fact that "pleasure" is not, in fact, a single-variable thing at all). Then, it would also have to believe that human pleasure is entirely consistent with forcing a human to submit to a dopamine drip against the most violent, screaming protestations that this was not wanted. The only way that the AI could take this attitude to the concept of human pleasure would be to change the concept in such a way that it flatly contradicts the usage prevailing in 99% of the human population (assuming that 99% of humans would scream "No!!").

So ... we are positing an artificial intelligence that is perfectly willing to take at least one existing concept and modify it to mean something that breaks that concept's connections to the rest of the conceptual network in the most drastic way possible. What part of

"maintaining the internal consistency of the knowledge base" don't we understand here, folks? What part of "from one logical contradiction, all false propositions can be proved" are we going to dump?

And yet we are to believe that this should be called a "higher" level of reasoning?

The Rot Spreads

If the AI is using this higher level of reasoning to come to conclusions about human happiness, it must also be using it in all of its other attempts to understand the world and deal with the threats that it faces. After all, it would make no sense to suggest (would it?) that the AI could commit that kind of concept-butchery in one circumstance, but in all other circumstances never repeat the mistake and only come to perfectly reasonable, safe and consistent conclusions.

But if the same thing is happening all the time in the life of this AI, who knows where its reasoning mechanism will take it? Given the task of, say, learning all about physics (so it could get enough knowledge to invent things that would make it more powerful than us), it might decide that "learn all about physics" is the same thing as "solve the exercises in the back of the physics book by copying them blindly from the internet". Or, when given the problem of learning how to control a robot arm to do really subtle movements, it might decide that the optimal strategy was to build a telepresence connection and out-source the robot-arm control to a human in India (or wherever people are outsourcing jobs to in the future!).

These two examples don't even scratch the surface. There is no limit to the extent of the AI's bizarre reasoning patterns if we allow the Drip-Feed-Equals-Happiness reasoning pattern to count as a "higher" form of reasoning.

But why don't we try to be as generous as possible and suppose that there might be a reason why it would commit bizarre acts of reasoning in the domain of human satisfaction, but at the same time never commit those bizarre acts when trying to make itself superintelligent. In that case, where are the proofs? Where is the theory-of-AI argument that explains why the AI will never disrupt its own path to superintelligence by committing trillions of similar acts of concept-butchery?

Or, put another way: why would anyone be tempted to describe this as a "higher level of reasoning" in the first place? If I were to suggest that the ramblings of a schizoid human with an IQ of 10 should count as a "higher level of reasoning", would my claim be any more or less reasonable than the suggestion that the AI is exhibiting a superior form of reasoning? In both of these cases the simplest conclusion, given the observed behavior, would be that neither of these individuals is going to be smart enough to do rocket science.

Inevitable Friendliness?

One last objection: "This seems to validate the idea of a friendly god-in-a-box that could

never do anything we disagreed with."

The argument I have presented targets only the wild inconsistency in a certain line of reasoning. I am attempting to eradicate something that looks suspiciously like a "have-your-cake-and-eat-it-too" argument: the idea that an AI could be so powerful that it was an existential threat, but at the same time so irrational that its understanding of the world could never have caused it to become superintelligent in the first place.

Notice that that is not the same thing as making the positive claim that every AI would be "a friendly god-in-a-box that could never do anything we disagreed with". The latter claim requires a good deal more argument, more or less distinct from the internal inconsistency that I am trying to bring to everyone's attention.

So the idea that AI might never disagree with us is an argument for another day. Let's keep them separate.

Conclusion

No, none of what I have just written is intended as an argument for complacency. There are plenty of issues and threats that need to be understood in depth, and there are several large mountains of debate still to be had.

But one thing that is a complete waste of time, tantamount to hysterical fearmongering, is to perpetually talk about a scenario that is riddled with internal logical inconsistency.

Worse, I have to say that in my opinion it counts as borderline fraud when organizations like the Singularity Institute try to sell that specious argument while asking for donations, and while at the same time dismissing the internal logical inconsistency with a scornful wave of the hand.

Ask for donations by all means. Suggest strategies for dealing with real threats, by all means. Study the threats with an unjaundiced eye that takes in all the possible ways to design an intelligent system, by all means. But don't try to fatten your coffers by suggesting that even the most sincerely friendly superintelligence might tile the universe with smiley faces, or kill every human as a way to minimize unhappiness, or put us all on a dopamine drip.

I can't resist the temptation to close on a humorous note, with an excerpt from Marvin's encounter with the Frogstar Scout robot class D. Marvin, here, is the real superintelligence, and his closing comment nicely captures my feelings about the concept of a dumb AI.

**********
[From Douglas Adams' Restaurant at the End of the Universe.]
Marvin looked pitifully small as the gigantic black tank rolled to a halt in front of him.
"Out of my way little robot," growled the tank.
"I'm afraid," said Marvin, "that I've been left here to stop you."

"You? Stop me?" roared the tank. "Go on!"
"No, really I have," said Marvin simply.
"What are you armed with?" roared the tank in disbelief.
"Guess," said Marvin.
"Errmmm ..." said the machine, vibrating with unaccustomed thought, "laser beams?"
Marvin shook his head solemnly.
"No," muttered the machine in its deep guttural rumble, "Too obvious. Anti-matter ray?" it hazarded.
"Far too obvious," admonished Marvin.
"Yes," grumbled the machine, somewhat abashed, "Er ... how about an electron ram?"
This was new to Marvin. "What's that?" he said.
"One of these," said the machine with enthusiasm. From its turret emerged a sharp prong which spat a single lethal blaze of light. Behind Marvin a wall roared and collapsed as a heap of dust. The dust billowed briefly, then settled.
"No," said Marvin, "not one of those."
"Good though, isn't it?"
"Very good," agreed Marvin.
"I know," said the Frogstar battle machine, after another moment's consideration, "you must have one of those new Xanthic Re-Structron Destabilized Zenon Emitters!"
"Nice, aren't they?" said Marvin.
"That's what you've got?" said the machine in considerable awe.
"No," said Marvin.
"Oh," said the machine, disappointed, "then it must be ..."
"You're thinking along the wrong lines," said Marvin, "You're failing to take into account something fairly basic in the relationship between men and robots."
"Er, I know," said the battle machine, "is it ..." it tailed off into thought again.
"Just think," urged Marvin, "they left me, an ordinary, menial robot, to stop you, a gigantic heavy-duty battle machine, whilst they ran off to save themselves. What do you think they would leave me with?"
"Oooh, er," muttered the machine in alarm, "something pretty damn devastating I should expect."
"Expect!" said Marvin, "oh yes, expect. I'll tell you what they gave me to protect myself with shall I?"
"Yes, alright," said the battle machine, bracing itself.
"Nothing," said Marvin.
There was a dangerous pause. "Nothing?" roared the battle machine.
"Nothing at all," intoned Marvin dismally, "not an electronic sausage."
The machine heaved about with fury. "Well, doesn't that just take the biscuit!" it roared, "Nothing, eh? Just don't think, do they?"
"And me," said Marvin in a soft low voice, "with this terrible pain in all the diodes down my left side."
"Makes you spit, doesn't it?"
"Yes," agreed Marvin with feeling.
"Hell that makes me angry," bellowed the machine, "think I'll smash that wall down!" The electron ram stabbed out another searing blaze of light and took out the wall next to the machine.

"How do you think I feel?" said Marvin bitterly.
"Just ran off and left you, did they?" the machine thundered.
"Yes," said Marvin.
"I think I'll shoot down their bloody ceiling as well!" raged the tank. It took out the ceiling of the bridge.
"That's very impressive," murmured Marvin.
"You ain't seeing nothing yet," promised the machine, "I can take out this floor too, no trouble!" It took out the floor, too. "Hell's bells!" the machine roared as it plummeted fifteen storeys and smashed itself to bits on the ground below.
"What a depressingly stupid machine," said Marvin and trudged away.

---

**Richard Loosemore** is a professor in the Department of Mathematical and Physical Sciences at Wells College, Aurora, NY, USA. He graduated from University College London, and his background includes work in physics, artificial intelligence, cognitive science, software engineering, philosophy, parapsychology and archaeology.

---

Newsletter: http://ieet.org/mailman/listinfo/ieet-announce

Contact: Executive Director, Dr. James J. Hughes,
IEET, 56 Daleville School Rd. Willington CT 06279 USA
Email: director@ieet.org
phone: 860-428-1837