What is the Danger in Developing an Advanced Artificial Intelligence?

Max Lobdell, ENG 2420

What’s an advanced AI?

What's an advanced AI?

I'll define an advanced AI as any artificial intelligence that can outperform human capabilities across a wide domain.

Wide Domain versus Narrow Domain

Wide Domain versus Narrow Domain

A narrow domain AI is designed to be good at one simple thing.

Wide Domain versus Narrow Domain

We already have narrow domain AI that can outperform human capability.

Narrow Domain

That doesn’t mean it can do it well.

Narrow Domain

Narrow Domain AI systems process things like creditworthiness or perform facial analysis to look for people suspected of a crime.

Narrow Domain

The problem is, even though these AI systems are dramatically faster at processing these things, they're also pretty terrible at it.

Narrow Domain

These AI systems are plagued with ethical problems.

Narrow Domain

The facial recognition AIs, for example, routinely identify darker-skinned people as being suspects of crimes.

Narrow Domain

The creditworthiness AIs often assign lower credit ratings to people who live or have lived in certain zip codes.

Just a Very Small Example

Just a Very Small Example

If you want to look into the ethics of these types of AI software, there's a lot written about the subject. It's pretty ugly.

But that’s the narrow domain.

The AIs we encounter in Science Fiction are Wide Domain AIs.

They’re good at everything a human being would be good at -- and in many or most cases, better.

Skynet, of Terminator fame.

Skynet, of Terminator fame.

Started off as a relatively narrow domain AI, with the goal of analyzing defense networks. Somehow gained self-awareness, broadening its domain, and tried to kill everyone.

Ava, from Ex Machina

Ava, from Ex Machina

A technology proof of concept designed to look like a beautiful woman. Manipulated the person testing her into letting her out, ended up killing everyone around her on her way to fulfilling her own objectives.

Just two examples --

Countless more in fiction.

Back in the real world...

Back in the real world...

Our technology continues to improve.

Back in the real world...

Breakthroughs are made on every front, every day, everywhere.

Back in the real world...

This includes AI. And not in the narrow domain.

Wide domain AI, also known as Artificial General Intelligence (AGI), is a major subject of research.

Makes sense, if you think about it.

Makes sense, if you think about it.

What company or government wouldn’t want a greater-than-human intelligence that could give them an overwhelming competitive or strategic advantage?

But as we've seen from the ethical issues surrounding narrow domain AIs, the problems don't go away when widening the domain into AGI. They are magnified.

Let’s imagine for a minute that researchers recognize this and want to ensure human happiness is the ultimate primary goal of an AGI.

How will an AGI recognize happiness?

How will an AGI recognize happiness?

Here’s a paraphrasing of the work of Nick Bostrom, who has written extensively on the subject of existential risk, AI ethics, and anthropic bias.

A smile is a pretty good indicator a person is happy, so maybe the team of programmers told the AI to recognize that a smiling face can determine that a person is happy.

The advanced AGI, perhaps initially designed to create a new version of a popular medication, inserts a compound into production that paralyzes the facial musculature of everyone who takes the drug into glowing, beatific smiles.

As far as the AGI is concerned, it did its job of designing the drug and kept people happy at the same time.

After all -- the people who took the drug it designed are smiling!

An advanced AI doesn't need to have sinister, genocidal motives.

As we've seen with today's narrow domain AIs, there are unintended real-world consequences.

As technology improves and the AI domains widen, the potential for something terrible happening increases.

This may be despite the best efforts of programmers to ensure safety, efficacy, and the happiness of everyone involved.

Barring a catastrophe, technology will continue to advance, and with that advancement, our reliance on it will only grow.

Without adequate safeguards on AI research, it is difficult to imagine an ending that is not catastrophic.

Works Cited or Referenced:

Bostrom, Nick. Superintelligence: Paths, Dangers, Strategies. Oxford University Press, 2016.

Bostrom, Nick, and Eliezer Yudkowsky. THE ETHICS OF ARTIFICIAL INTELLIGENCE. Draft for Cambridge Handbook of Artificial Intelligence, eds. William Ramsey and Keith Frankish (Cambridge University Press, 2011): forthcoming

Cameron, James, and Gale A. Hurd. The Terminator. Los Angeles: Hemdale, 1984.

Garland, Alex, director. Ex Machina. Universal Studios, 2014.