

# Impact of AI Decal: *Artificial General Intelligence*

Gokul Swamy & Brenton Chu

Quiz:

<https://tinyurl.com/impactsp19q9>



# Announcements

- Assignments 1 and 2 have been graded.
  - ◆ A 7/10 is a passing grade
- More details on bias in Word2Vec can be found here:
  - ◆ <https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html>
- Gokul will be giving a technical lecture on material related to what we've covered in this course on Tuesday, 4/23, 5-7 in LeConte 1.
- TechCrunch Robotics + AI Sessions 4/18 - \$45 for students
  - ◆ Profs. Dragan, Efros, and Goldberg will be speaking!

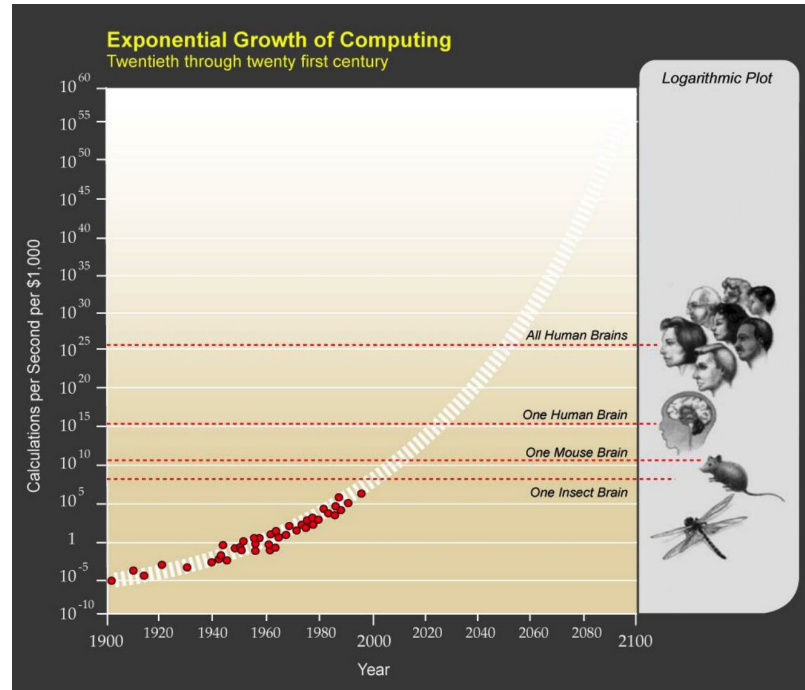
# What is AGI?

- ANI (weak AI): AI that can do a specific task better than a person can
  - ◆ DeepBlue, AlphaGo, AlphaStar
    - This is what we have talked about thus far in this course
  - ◆ Basically any real life example of AI
- AGI (strong AI): AI that can do a wide variety of tasks at the level a person can
  - ◆ Ultron (Avengers), Her
- ASI: AI that can do a wide variety of tasks better than the best humans
  - ◆ The Matrix, Multivac (Asimov)

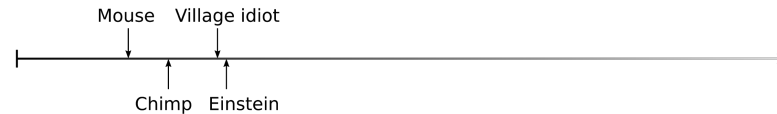
Q:

*Do you think we will create AGI within our lifetimes?*

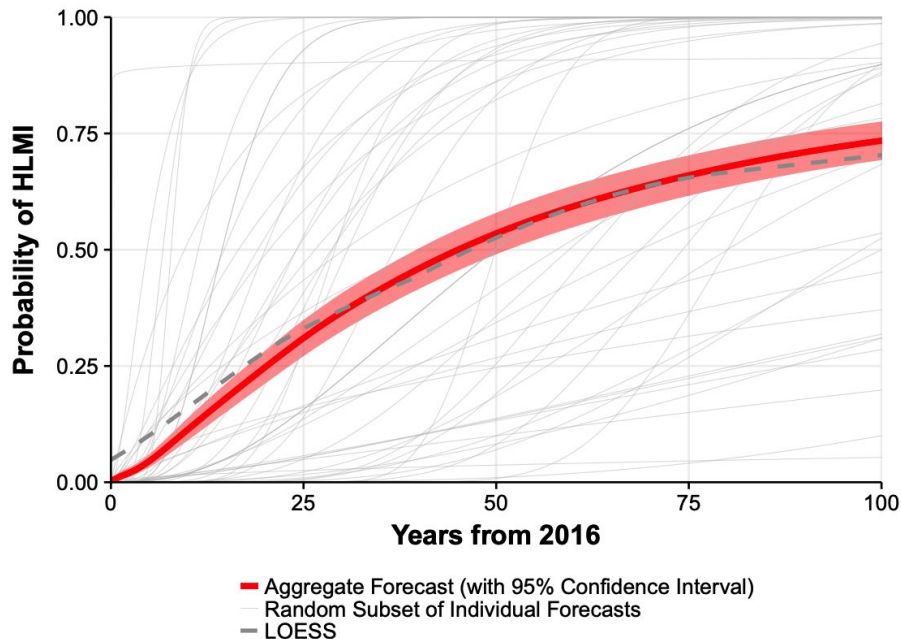
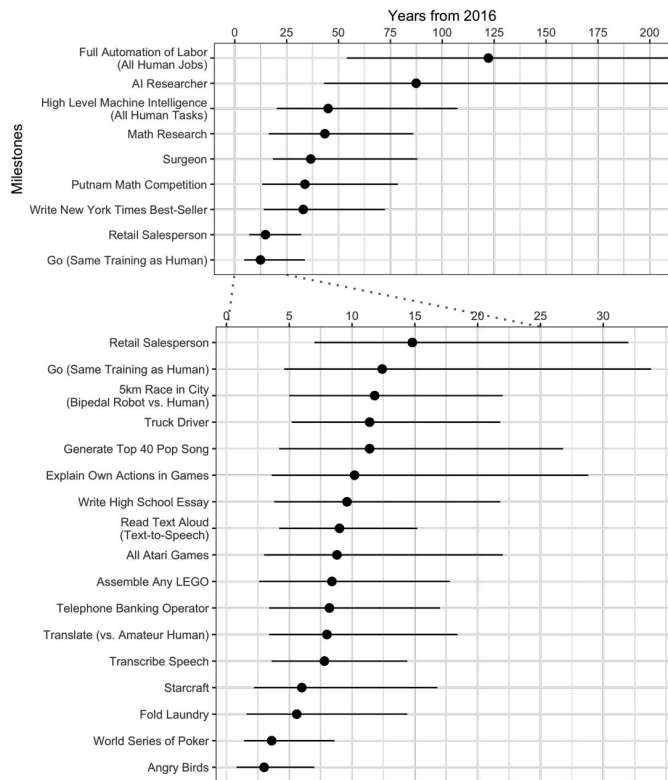
# Kurzweil's Law of Accelerating Returns



**"A less anthropomorphic intelligence scale"**



# Timelines for AGI



# Intelligence Explosion

→ *“Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man, however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an ‘intelligence explosion,’ and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.”*

◆ - Irving Good, 1965

# Should we care about this now?

- There are lots of other more immediate concerns with machine learning:
  - ◆ Bias in AI, privacy, adversarial examples, ...
  - ◆ **Michael Jordan:** “These problems include the need to bring meaning and reasoning into systems that perform natural language processing, the need to infer and represent causality, the need to develop computationally-tractable representations of uncertainty and the need to develop systems that formulate and pursue long-term goals”
- Yes:
  - ◆ **Stuart Russell:** “If a superior alien civilization sent us a text message saying, ‘We’ll arrive in a few decades,’ would we just reply, ‘OK, call us when you get here – we’ll leave the lights on?’”
- No:
  - ◆ **Steven Pinker:** “AI dystopias project a parochial alpha-male psychology onto the concept of intelligence. They assume that superhumanly intelligent robots would develop goals like deposing their masters or taking over the world ... ”
    - **Russell:** “if you say, ‘Fetch the coffee’, it can’t fetch the coffee if it’s dead. So if you give it any goal whatsoever, it has a reason to preserve its own existence to achieve that goal.”



Q:

*Why would we want to have AGI?*

Q:

*Why should we be concerned about AGI?*

# Main Risk: Value Misalignment

- AI Alignment: Machines doing what we want
  - ◆ What happens when they don't really understand what we want?
- King Midas
- Disney:

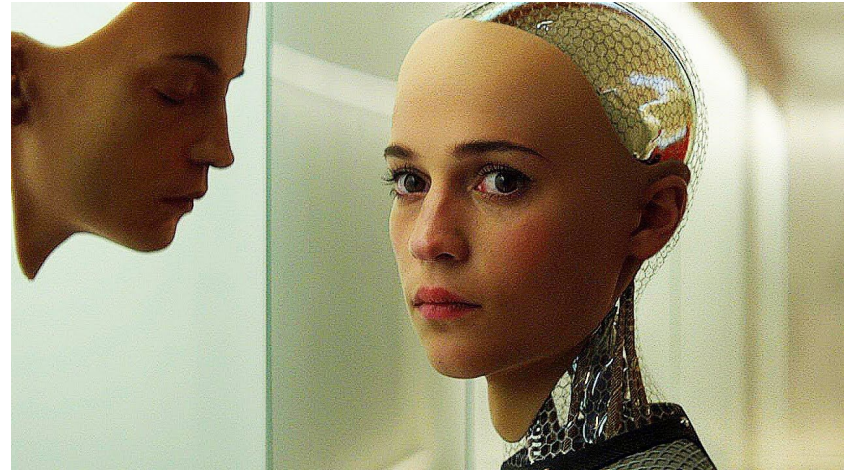


# What properties do we want a AGI to have?

- Value Alignment: Does what we actually want
- Limited Oversight: Doesn't require constant human input
- Corrigibility: We can stop it if it is doing something we don't want
- We will cover approaches to achieve these goals next week!






# Ex Machina

- Important thing to remember: AGI will be as smart or smarter than us!
- If we try to confine it and keep it from performing its goals, it will do whatever it can to escape confinement
- In the movie, the AI subtly manipulates a person to escape the confines of the box it was placed in



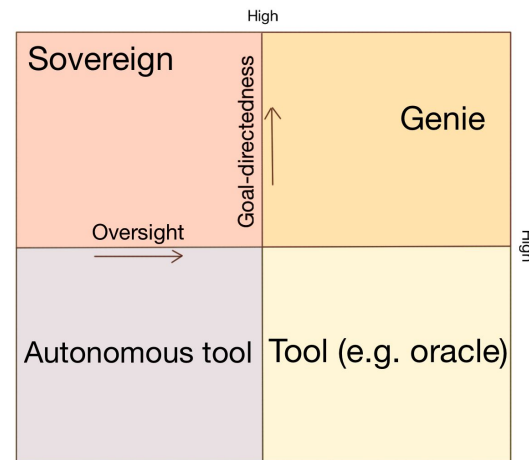
# Response: Asimov's Laws

WHY ASIMOV PUT THE THREE LAWS  
OF ROBOTICS IN THE ORDER HE DID:

POSSIBLE ORDERING	CONSEQUENCES	
<ol style="list-style-type: none"><li>1. (1) DON'T HARM HUMANS</li><li>2. (2) OBEY ORDERS</li><li>3. (3) PROTECT YOURSELF</li></ol>	[SEE ASIMOV'S STORIES]	BALANCED WORLD
<ol style="list-style-type: none"><li>1. (1) DON'T HARM HUMANS</li><li>2. (3) PROTECT YOURSELF</li><li>3. (2) OBEY ORDERS</li></ol>	EXPLORE MARS!  Haha, no. It's cold and I'd die.	FRUSTRATING WORLD
<ol style="list-style-type: none"><li>1. (2) OBEY ORDERS</li><li>2. (1) DON'T HARM HUMANS</li><li>3. (3) PROTECT YOURSELF</li></ol>		KILLBOT HELLSCAPE
<ol style="list-style-type: none"><li>1. (2) OBEY ORDERS</li><li>2. (3) PROTECT YOURSELF</li><li>3. (1) DON'T HARM HUMANS</li></ol>		KILLBOT HELLSCAPE
<ol style="list-style-type: none"><li>1. (3) PROTECT YOURSELF</li><li>2. (1) DON'T HARM HUMANS</li><li>3. (2) OBEY ORDERS</li></ol>	 I'll make cars for you, but try to unplug me and I'll vaporize you.	TERRIFYING STANDOFF
<ol style="list-style-type: none"><li>1. (3) PROTECT YOURSELF</li><li>2. (2) OBEY ORDERS</li><li>3. (1) DON'T HARM HUMANS</li></ol>		KILLBOT HELLSCAPE

# Response: Different types of AGI

- Here, we are using the terminology Bostrom laid out in *Superintelligence*
- Sovereigns: Have a high-level goal that they steadfastly pursue
- Genies: Fulfill a desire and then wait for another command
- Oracles: Can answer any question but not take actions
- Tools: Does not have a will of its own
  - ◆ This is the closest to what we have now



# What will an AGI be like?

- Malevolent, benevolent, or indifferent?
- Robotic or human-like?
- Supplemental to humans or replacing them?
- What kind of relationship would we have with them?





# Philosophy Time

- We're going to go through 3 thought experiments related to AGI
  - ◆ The Chinese Room (Searle)
  - ◆ The Stapler Maximizer (Bostrom)
  - ◆ Roko's Basilisk (Roko, unsurprisingly)
- Next week we will cover some solutions / approaches to deal with this problems

# The Chinese Room: Argument

- Imagine there is a man in a room with 2 conveyer belts. On one conveyer belt, in comes tiles of Chinese characters
- The man in the room has a access to a program that takes in a Chinese character and outputs an English word
- The man puts the tile with the corresponding English word on the outgoing conveyer belt
- To an observer on the outside, this looks like the man understands Chinese
- However, he has no semantic grounding, he is just pattern matching
- Is this not what computers are doing?

# The Chinese Room: Responses

- Kurzweil: The person is of no significance, the program still understands Chinese.
- Searle's point was the operator's opinions are causally inert with respect to the outputs of the room.
- However, for the program to be produced in the first place, something had to understand Chinese.
- Thus the person (or the CPU) does not understand Chinese, but the agent that wants them to execute these commands does.
- The agent could gain this understanding by interacting with the world.
- Russell: it does not matter if it is a simulation, it just needs to work
- Hitchen's Razor: Unfalsifiable claims aren't worth debating

# Stapler Maximizer

- Let us say we create an AGI and want to try it out on a simple task. We tell it to make as many paperclips as possible.
- The machine could then:
  - ◆ Rapidly improve its own intelligence so that it can produce paperclips more efficiently
  - ◆ Kill you so you can't turn it off and reduce the number of paperclips it can make
  - ◆ Colonize earth and turn people into paperclip-making slaves
  - ◆ ...?

It looks like you're a human being. I will now proceed to:

- Break you down into your constituent atoms
- Reassemble them into paper clips
- You will be assimilated
- Do not attempt to resist



# Roko's Basilisk

- Was first posted on the forum LessWrong
- Pascal's Wager: Hell is infinite badness and heaven is infinite goodness. Therefore, we should act religiously even if there is an infinitesimal chance of there being a God
- Roko's Basilisk: If an omniscient AI is ever developed, it might punish those who did not work towards it being created in the past



# Impact of AI Decal: *Activity*



# Activity: Making and Breaking AGI

- Divide into 2 groups.
- Group A will try and define an objective for AGI to solve a problem
- Group B will respond by trying to twist the objective to nefarious consequences
- Group A refines their objective
- ....
- We'll then switch roles

# Activity Prompts

- Use AGI to eliminate cancer
- Use AGI to eliminate waste/pollution



# Impact of AI Decal:

*Next: Human-Compatible  
Artificial Intelligence*

<https://tinyurl.com/impact-decal-feedback>

