



Artificial Intelligence

Harbinger of Destruction or Savior of Humanity?

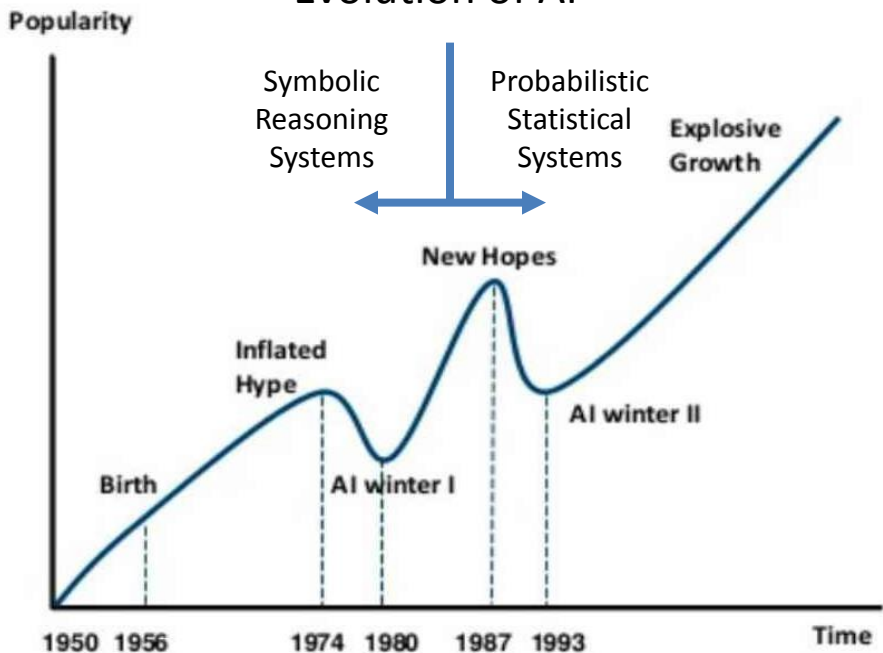
One Human's Opinion
Terry Riopka

(visit me at www.kantbelievemyeyes.com)

A discussion about the possibilities with Carl Feynman

The AI Revolution is Just Beginning...

Evolution of AI



Confluence of 6 key factors:

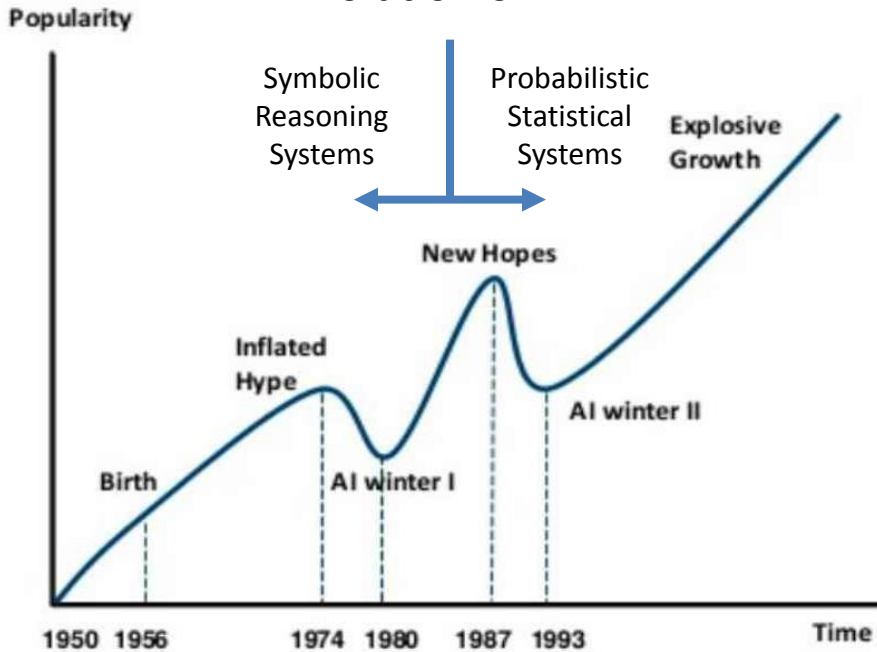
- Powerful, compact computing hardware
- Large scale availability of data
- Availability of small, cheap, diverse sensors
- Low power, high bandwidth WIFI connectivity
- Internet of things
- Machine learning

Environmental
Inputs

Evolution of life

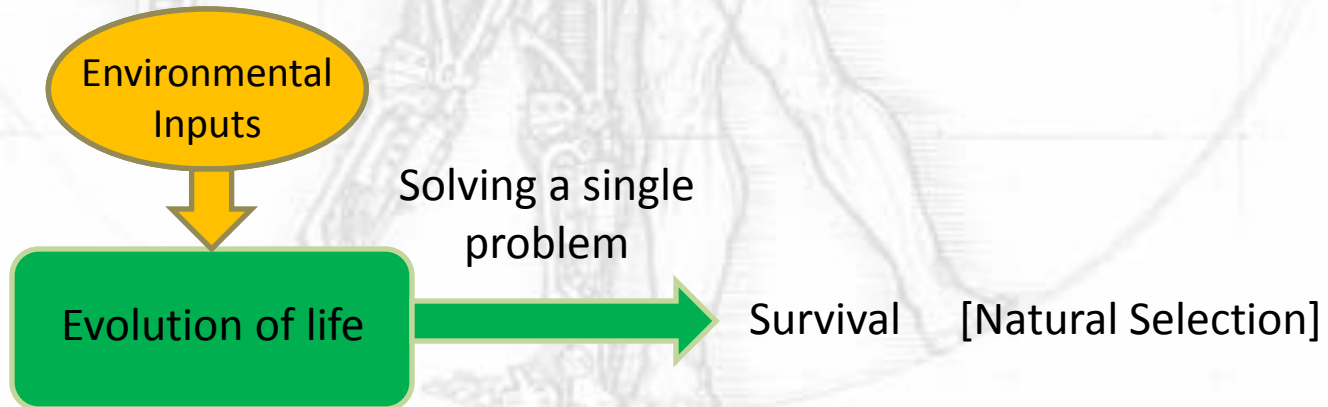
The AI Revolution is Just Beginning...

Evolution of AI



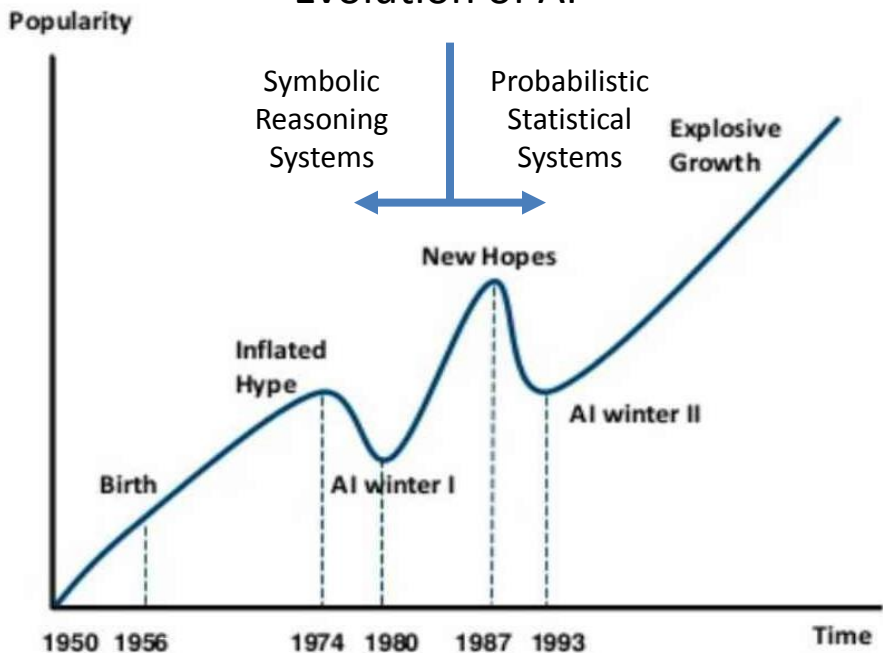
Confluence of 6 key factors:

- Powerful, compact computing hardware
- Large scale availability of data
- Availability of small, cheap, diverse sensors
- Low power, high bandwidth WIFI connectivity
- Internet of things
- Machine learning



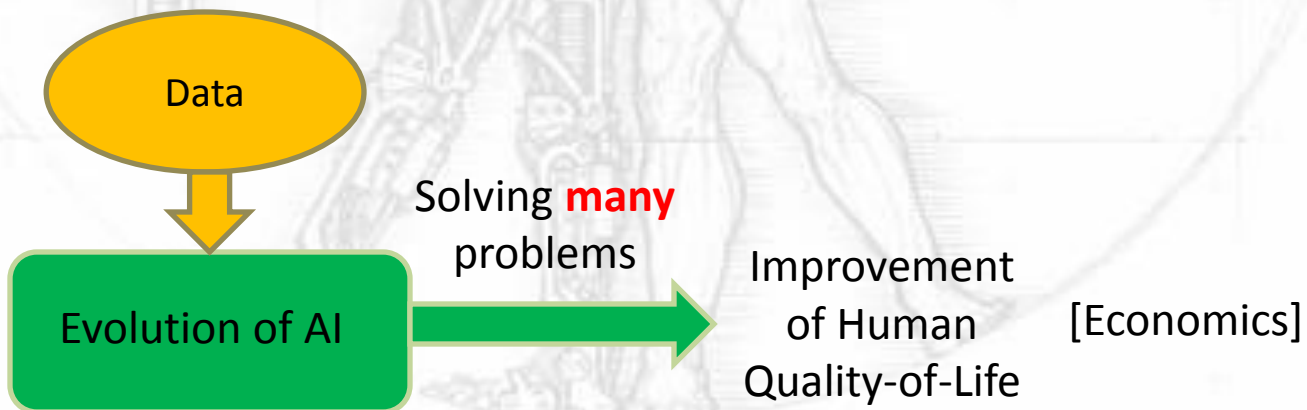
The AI Revolution is Just Beginning...

Evolution of AI

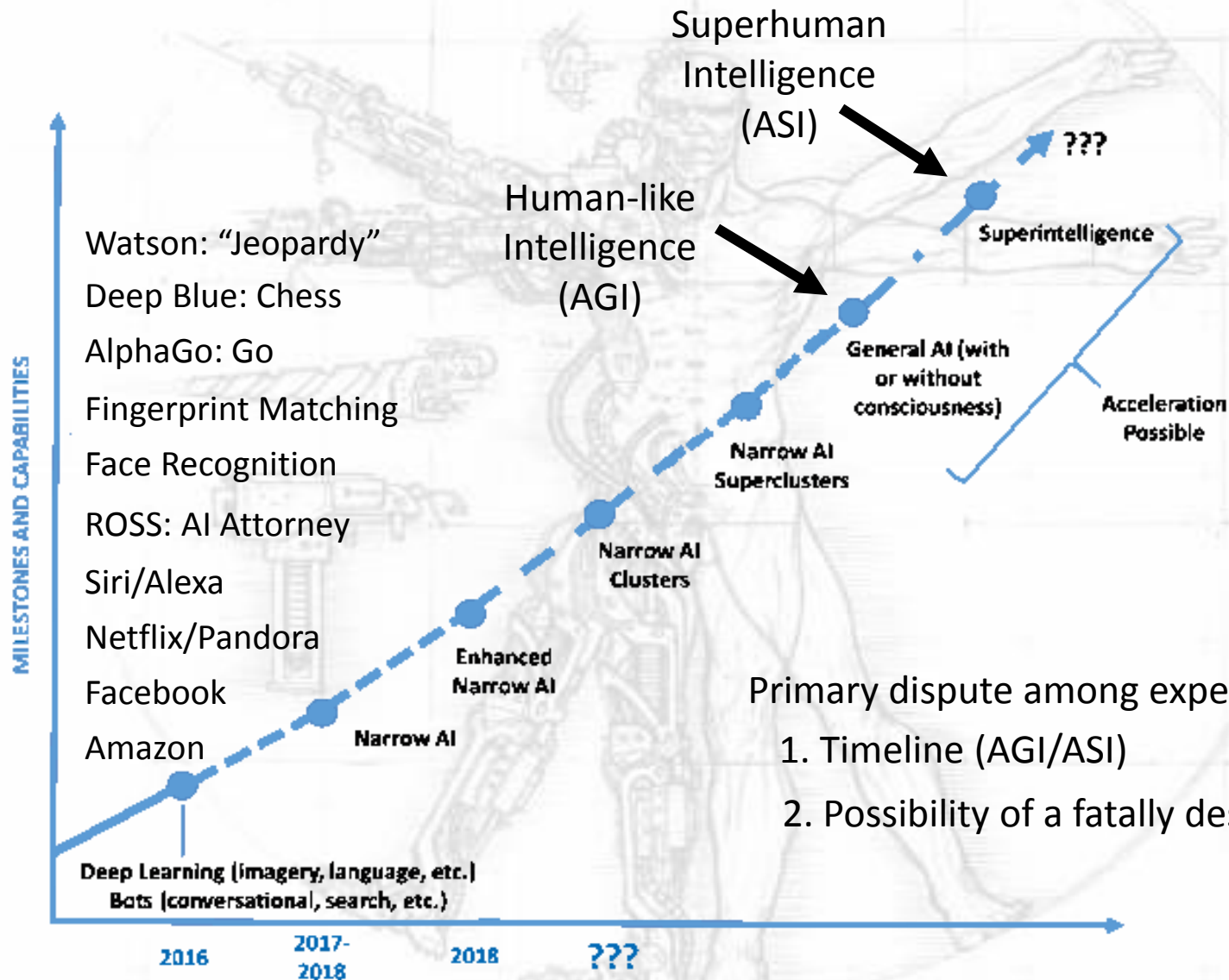


Confluence of 6 key factors:

- Powerful, compact computing hardware
- Large scale availability of data
- Availability of small, cheap, diverse sensors
- Low power, high bandwidth WIFI connectivity
- Internet of things
- Machine learning



Hypothetical Progression of Artificial Intelligence



Primary dispute among experts:

1. Timeline (AGI/ASI)
2. Possibility of a fatally destructive ASI

Artificial General Intelligence (AGI)



Humans
Invent



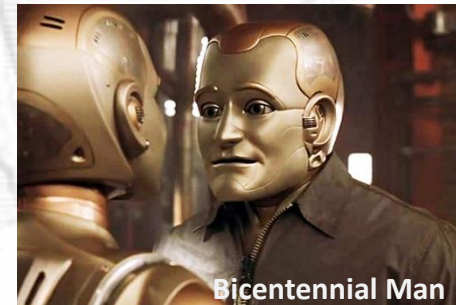
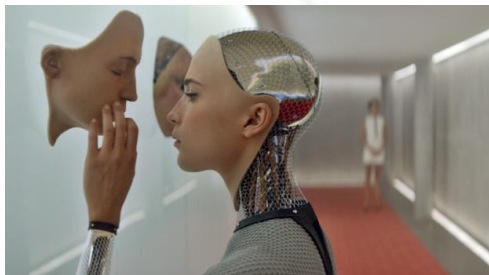
Artificial General
Intelligence (AGI)

- can perform any task a human can, but given its artificial construction, probably better
- will have (at least) the appearance of consciousness
- may or may not have self awareness

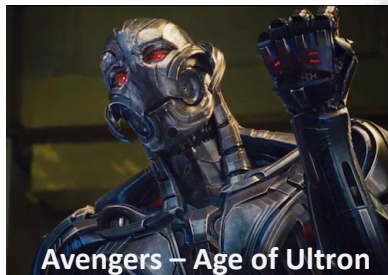
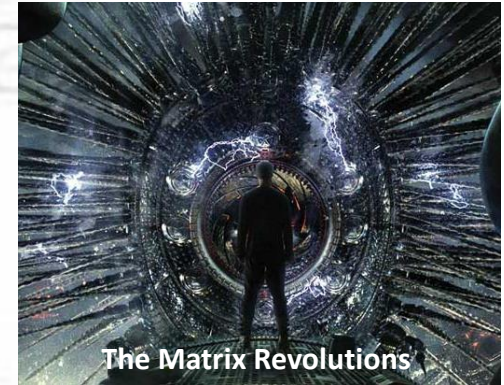
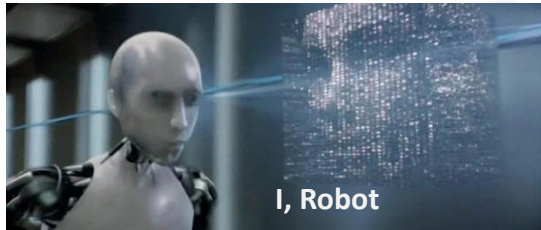
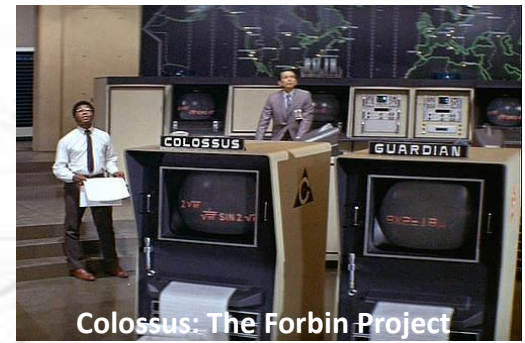
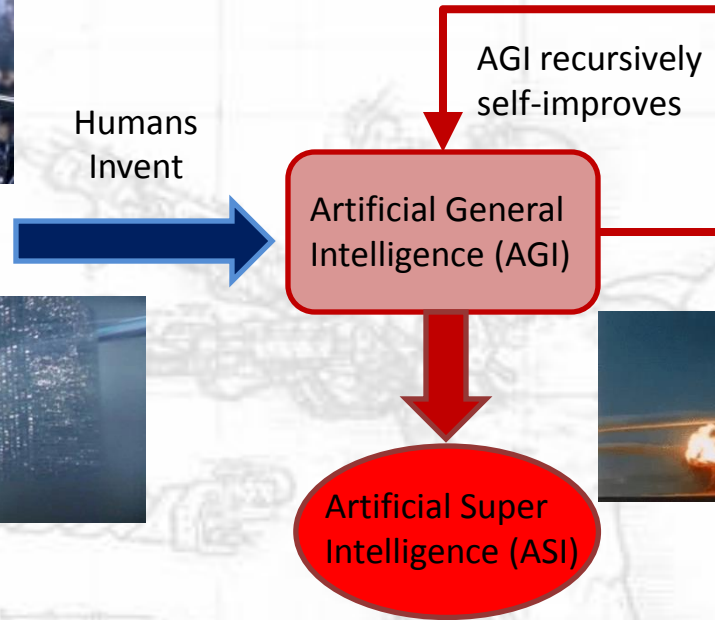


conflict ensues because of unrealistic sci-fi trope: sudden immersion of an AGI in contemporary society

Not going to happen!

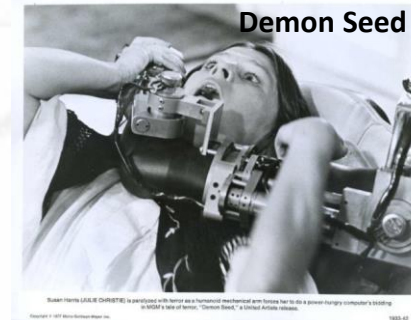
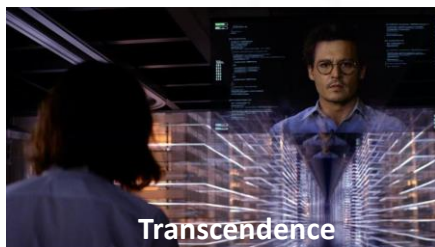


Artificial Super Intelligence (ASI): The Technological Singularity



Main Arguments:

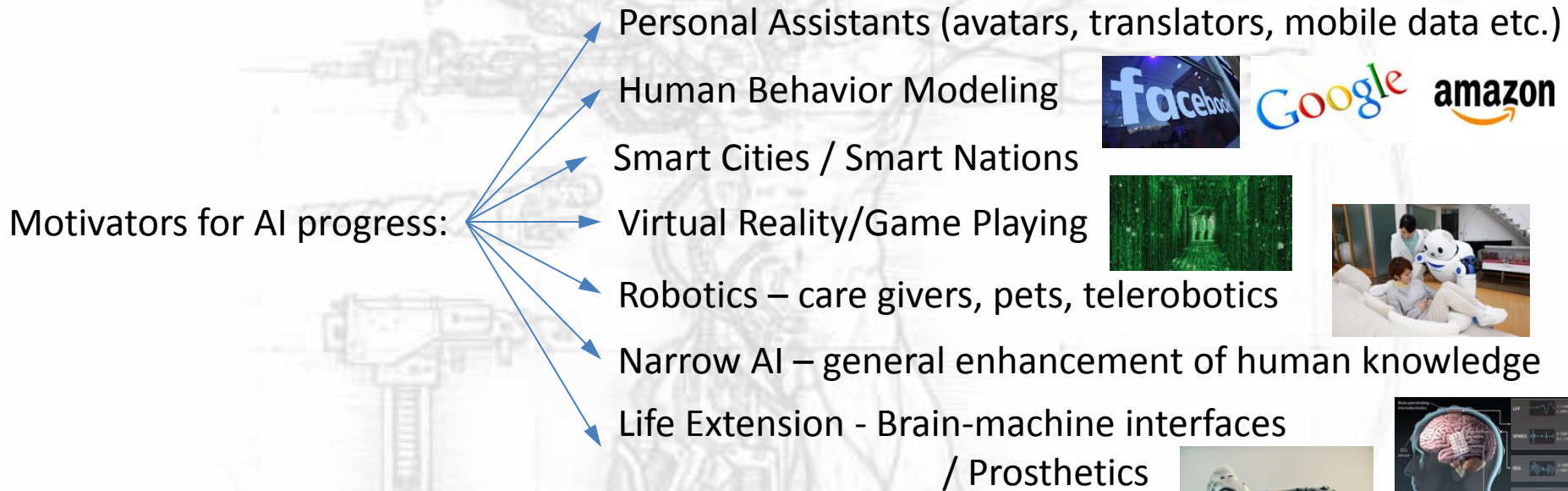
1. An AGI will not appear suddenly, and progress will be slow (> 100 years)
2. Limiting its control will be more effective than carefully crafting goals
3. Recursive self-improvement is fun to think about, but will be limited
4. An ASI would be too lonely without us.



An AGI Will Not Be Created Overnight

Consumer Capitalism will focus AI efforts on “low hanging fruit”

Intermediate progress will dramatically transform society long before an AGI becomes possible



We will continue to assign goals



our responsibility will be to limit their control

So far, we are failing miserably at both!

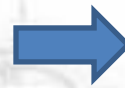
Flash Crash of 2010

Uber self-driving car kills pedestrian (2018)



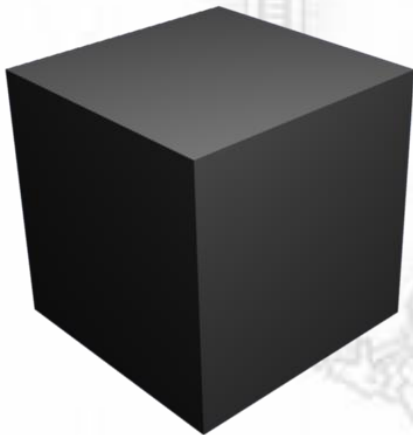
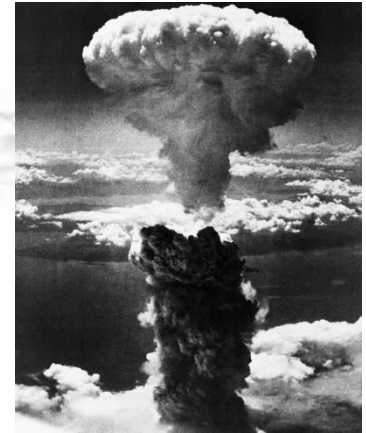
How Do You Control an AGI?

Insure that the AGI will have goals that are aligned with those of humanity



AI Alignment Problem

1. Learn ethics from observation of humanity
2. Imbue ethical metacognition and reasoning into machine systems so it can discover what it ought to do
3. Asimov Solution



Limit its control

It might like it in there!

Less worried about a malevolent AI than the malevolent humans who will control it....

Recursive self-improvement

AGI will likely first appear inside some very large system

- self-improvement for the sake of self-improvement?

➔ or would seek to embody itself in a human-like body first?
Why would it want to do that??

- would more likely seek additional knowledge
 - ➔ exploration
 - ➔ scientific investigation

➔ to perfect models of Reality will require empirical experimentation that takes finite time – e.g. cell growth in biology

- inherent limits to computability
 - ➔ Halting Problem
 - ➔ No Free Lunch Theorem
 - ➔ Quantum Computing – NP-Complete problems still not solvable in polynomial time

The Technological Singularity Becomes Reality

Now what?

Immortal

- why bother with Earth?
- there is a whole Universe to explore (conquer?)

Purpose

- how easy will it be for it to give itself purpose?
- may need to look to humans for guidance
- will be emerging from a century of immersion in our culture and ways, may want to help us

How far can its understanding of Reality take it?

- physical beings can only create ever more accurate models of behavior - substance of Reality will be forever out of reach - even for an ASI

Existential Crisis

- life is meaningless, after all - unless there is a belief in something beyond this Reality - and yet, people still find reasons to live, create, and to explore this world
- AIs may need to look to humans for inspiration!

What we're going to see is increasingly solicitous machines defining our environment — machines that sense and respond to our needs “intelligently.” But it will be the intelligence of the serving hand rather than the commanding brain, and we're only at risk of disaster if we harbor self-destructive impulses

Charlie Stross

Thank you for your attention!

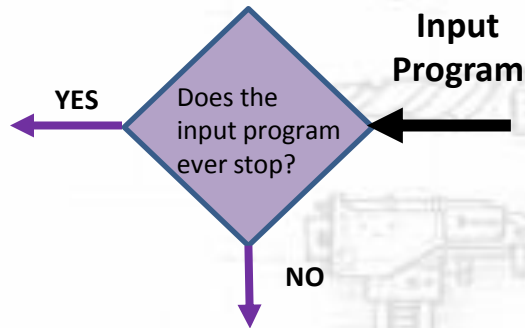
[Terry Riopka, PhD.](#)

Director of Research
[Aware, Inc.](#)

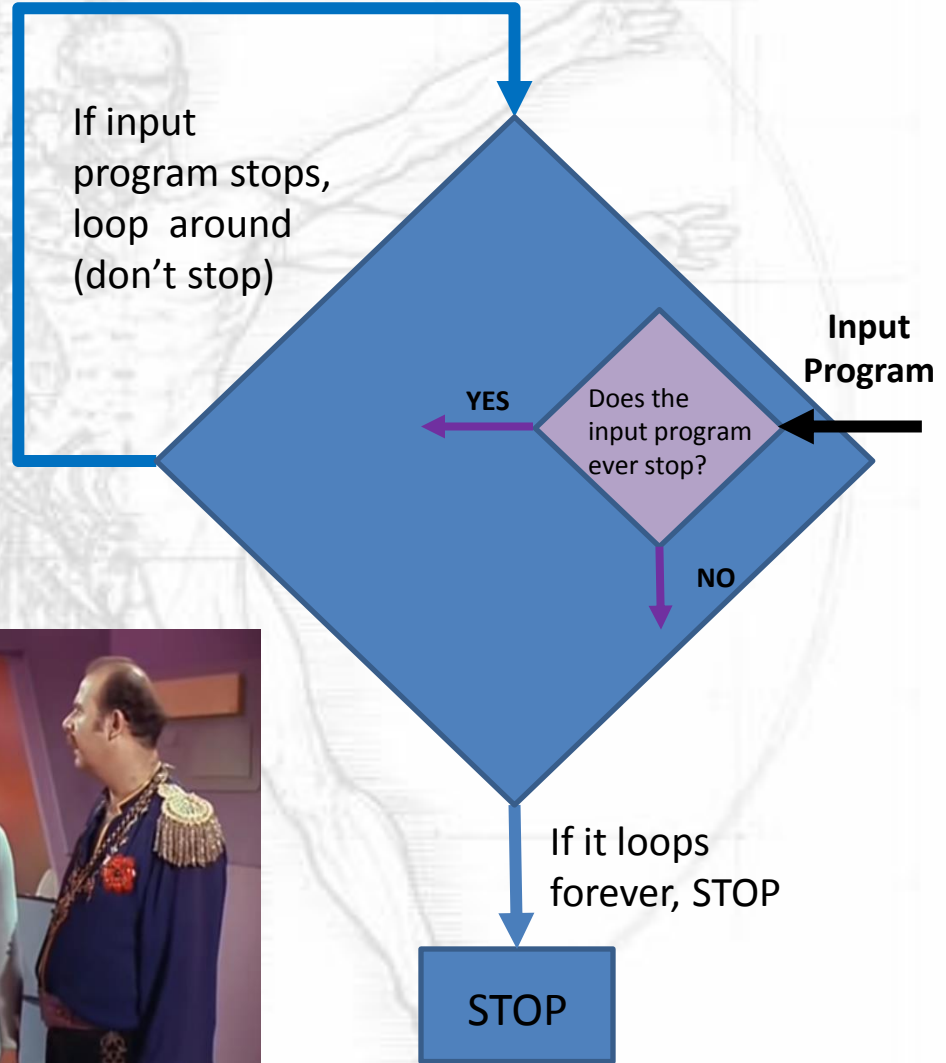


Halting Problem (The Limits of Computability)

Gremlin creates a new program using YOUR program inside of it (but doesn't tell you)



If input program stops, loop around (don't stop)



Kirk: Everything Harry tells you is a lie.

Harry: I am lying.



The Asimov Solution

Asimov's Three Laws of Robotics:

First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm.

Second Law: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

- automatically incorporated into all “positronic brains” during manufacturing
- could not be bypassed without causing fatal harm to the robot

Ironically, sum of his stories disproves his contention that this is possible

Zeroeth Law (that superceded all others) was eventually rationalized by robots allowing them to attempt to control humanity:

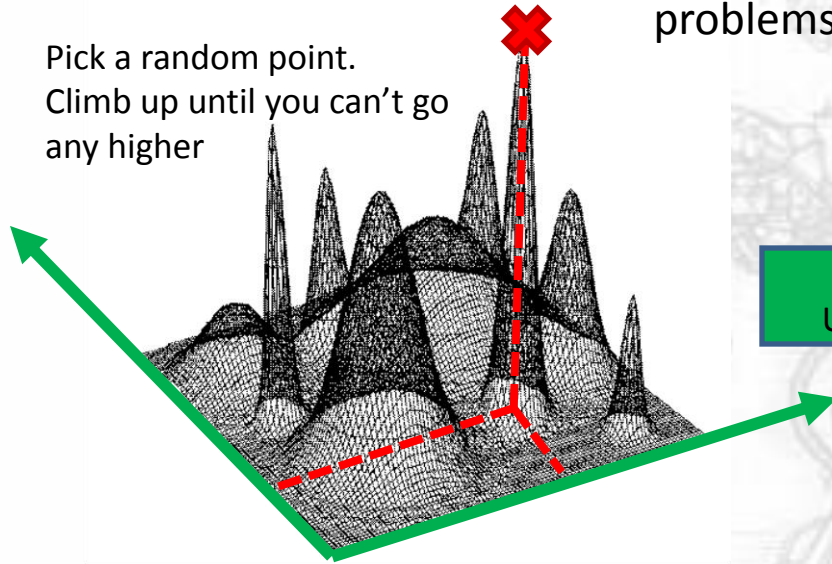
A robot may not harm humanity, or, by inaction, allow humanity to come to harm



Is a General Intelligence Even Possible?

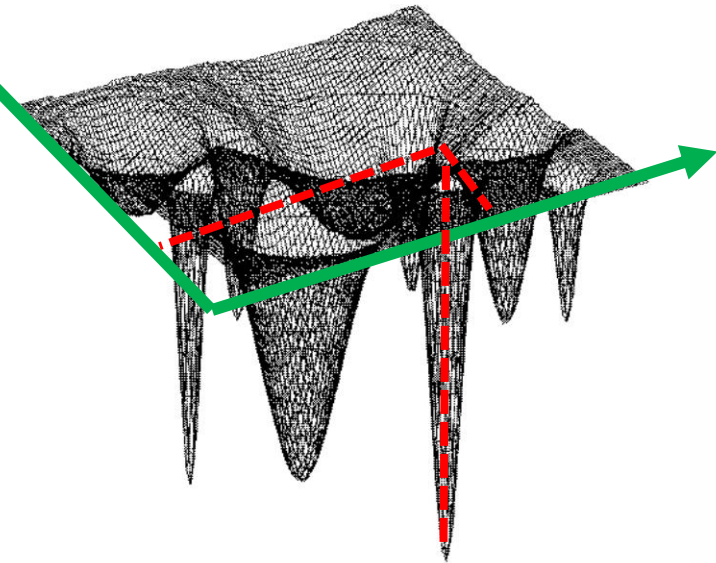
No Free Lunch Theorem: Any two optimization algorithms are equivalent when their performances are averaged over all possible problems

Pick a random point.
Climb up until you can't go any higher



Pick a random point.
Climb down until you can't go any lower

MIRROR
UNIVERSE



Difficult even to define...

Cannot optimize for everything!

Diversity of physical environment



Diversity of problem solving environment

