A Reality Check on AI

Luke Hutchison

PhD, MIT CSAIL; former senior research scientist, Google Machine Intelligence

FEATURING: MACHINE LEARNING ➤ DRIVERLESS CARS ➤ SUPERHUMANS

POPULAR SCIENCE

SPECIAL EDITION

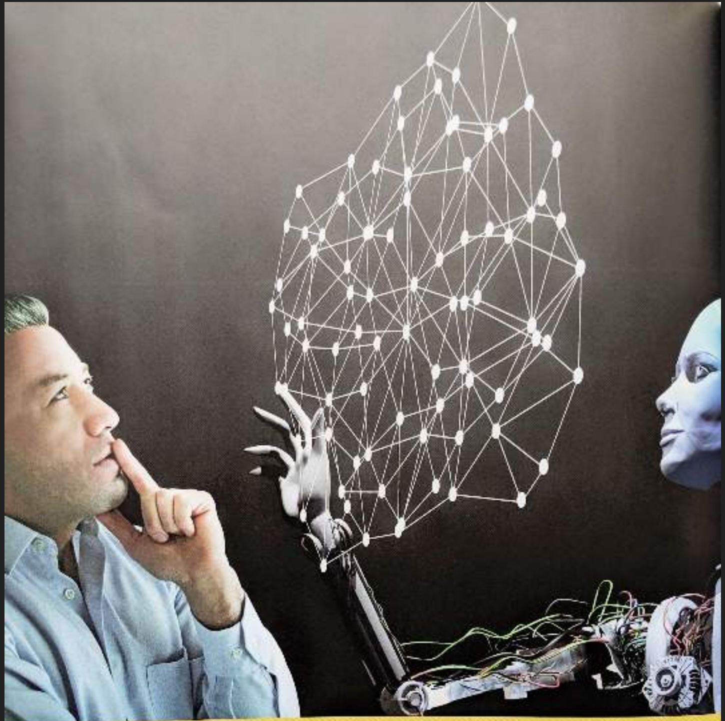
THE NEW ARTIFICIAL INTELLIGENCE

THE FUTURE OF NANOBOTS

DETECTING CANCER

KILLER ROBOTS?

A REISSUE OF A SPECIAL EDITION
DISPLAY UNTIL 10/4/19
$13.99US $16.99CAN
SPC SPECIALS



Scared of Artificial Intelligence? You Should Know It's Already Everywhere

From sandwich pics to real estate.

Google Has a List of A.I. Behaviors that Would Scare It Most

And it just might scare you, too.

Google is one of the companies at the forefront of robotics and artificial intelligence research, and being in that position means they have the most to worry about. The idea of a robot takeover may still be an abstract, science fictional concept to us, but Google has actually compiled a list of behaviors that would cause them great concern, for both efficiency and safety in the future.

Among the recurring themes is the possibility of deceit: that robots would learn to "accomplish goals" by hiding evidence that more work needs to be done. The researchers also laid out the possibility of smart robots avoiding humans to evade further assignments, disabling their own sensors to avoid finding tasks to do, discarding cellphones along with stray candy wrappers, or destroying furniture and other objects as a side effect of cleaning faster. Other disastrous risks include a robot testing more efficient mopping strategies—and deciding to try putting a wet mop into an electrical outlet.

The paper, "Concrete Problems in AI safety," was written in 2016 by members of Google Brain and

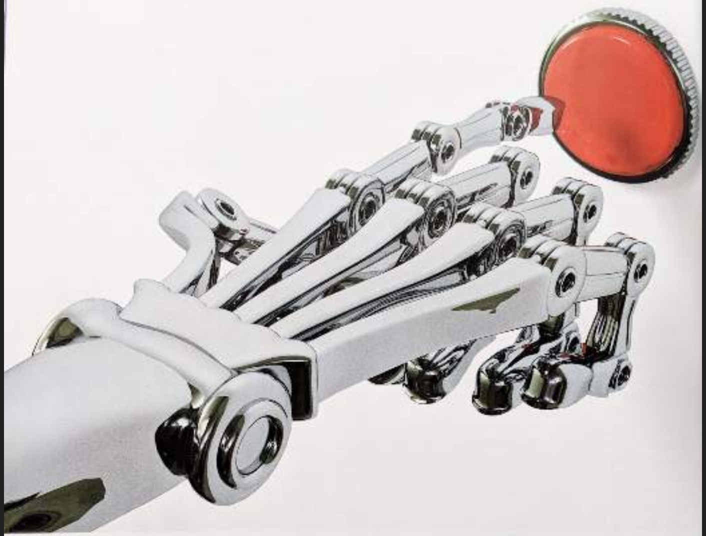
Who Will Driverless Cars Decide to Kill?

THEY FOUND THAT 75 PERCENT SUPPORTED SELF-SACRIFICE OF THE PASSENGER TO SAVE 10 PEOPLE, AND AROUND 50 PERCENT SUPPORTED SELF-SACRIFICE WHEN SAVING JUST ONE PERSON.

In 2016, researchers from the Toulouse School of Economics decided to see what the public would decide, and posed a series of questions to online survey-takers, including a situation where a car would either kill 10 people and save the driver, or save the group by swerving and killing the driver.

They found that more than 75 percent supported self-sacrifice of the passenger to save 10 people, and 50 percent supported self-sacrifice when saving just one person. However, respondents didn't actually think real cars would end up being programmed this way, and would probably save the passenger at all costs.

The questions were answered by paid participants on Amazon Mechanical Turk (compensated



Google Considers Making a Big Red Button to Stop Dangerous AI in an Emergency

In conjunction with AI doomsayer Nick Bostrom's Research Institution

As far as we know, artificial intelligence is the best way to automate vast complex tasks, like tagging millions of unique photos on Facebook or teaching robots to walk. At a breakneck pace, computer scientists and roboticists are getting better at crafting these algorithms, which means that now is the time to think about stopping AI's capacity to do wrong.

Google DeepMind, in conjunction with The Future of Humanity Institute, has released a study that determines how we would stop an artificially intelligent algorithm or robot if it were to go rogue. Their conclusion? A big red button.

The study points to earlier research conducted in 2013 where a game-playing algorithm realized that if it just paused Tetris, it would never lose.

How to Create Super-Intelligent Machines That Won't Kill Us

Preventing an Age of Ultron

In 2015's installment of the Marvel *Avengers* franchise, the artificial intelligence Ultron is hell-bent on exterminating humanity. In Ultron's own words, "I was designed to save the world," but the robot ultimately concludes that when it comes to humans, "there's only one path to peace: your extinction."

The advances that scientists are now making with artificial intelligence lead many to suggest—and fear—that we may be on the verge of creating artificial intelligences smarter than we are. If humanity does succeed in developing an artificial superintelligence, how might we prevent an Age of Ultron? That is exactly the kind of problem that Nick Bostrom, founding director of The Future of Humanity Institute at the University of Oxford, tackled in his 2014 book *Superintelligence: Paths, Dangers, Strategies*.

The fact that Ultron wants to save the world by eradicating humanity is what Bostrom might call "perverse instantiation"—an AI discovering some way of satisfying its final goal that violates the intentions of the programmers who defined the goal. For example, if one asks an AI to make a person smile, the computer might try manipulating facial nerves to paralyze the face into constantly smiling. If one then asks the machine to make us happy, the computer might then simply implant electrodes into the pleasure centers of our brains.

Should we be concerned?

“Open Letter on Artificial Intelligence” [2015]

(Elon Musk, Peter Norvig, Stuart Russell, Stephen Hawking, x150):

“...we could one day lose control of AI systems via the rise of superintelligences that do not act in accordance with human wishes – and that such powerful systems would threaten humanity. Are such dystopic outcomes possible? If so, how might these situations arise? ...What kind of investments in research should be made to better understand and to address the possibility of the rise of a dangerous superintelligence or the occurrence of an ‘intelligence explosion’?”

How soon could this happen?

How soon could this happen?

“In from three to eight years,
we will have a machine with
the general intelligence
of an average human being.”

[Who said this, and when?]

How soon could this happen?

“In from three to eight years,
we will have a machine with
the general intelligence
of an average human being.”

--Marvin Minsky, 1970 (in LIFE Magazine)

I worked with 300 of the best minds in AI at Google...

And honestly nobody even has a clue what intelligence is, or how to really build it.

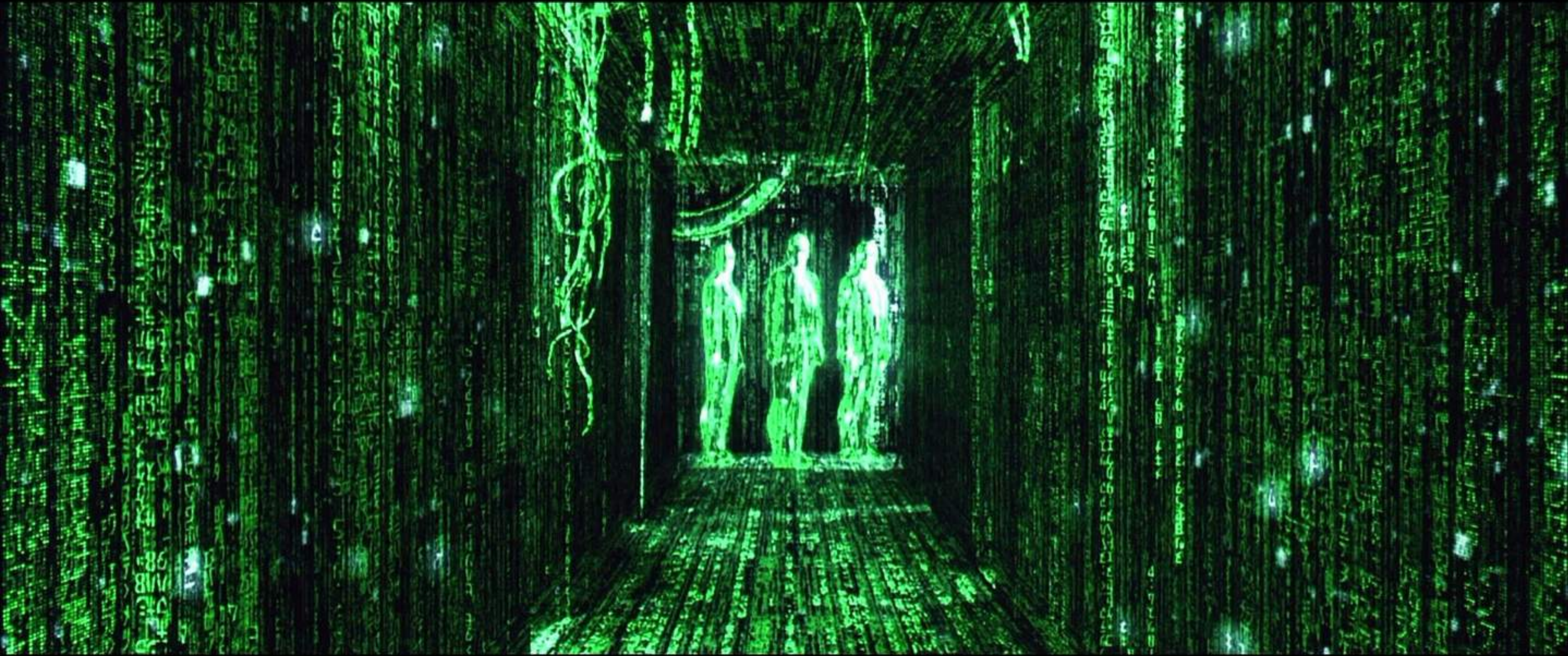
So I offer you a reality check on

AI hype

vs.

AI reality

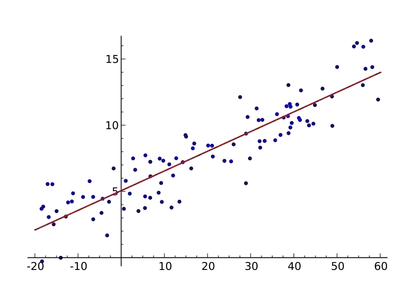
(“the red pill”)



What is AI -- really?

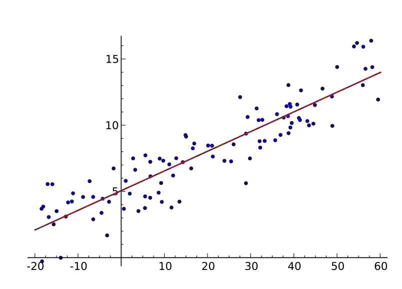
What is AI -- really?

- Supervised learning
  => Statistical regression

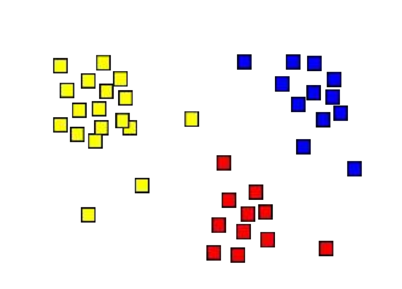


What is AI -- really?

- Supervised learning
  => Statistical regression

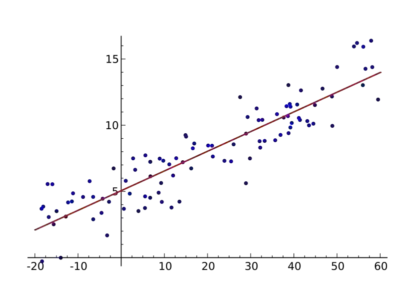


- Unsupervised learning
  => Clustering

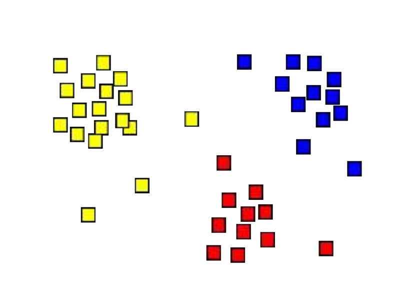


What is AI -- really?

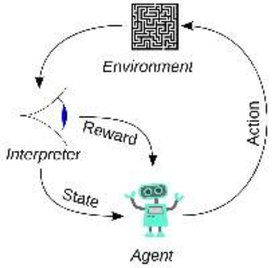
- Supervised learning
  => Statistical regression



- Unsupervised learning
  => Clustering

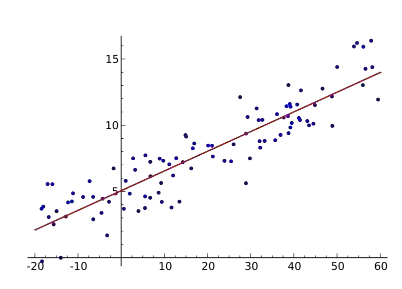


- Reinforcement learning
  => exploration / exploitation

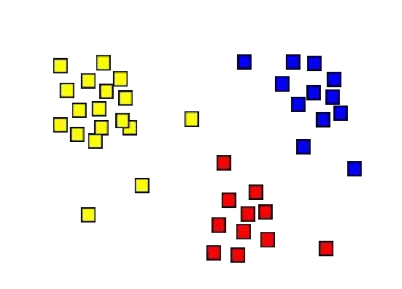


What is AI -- really?

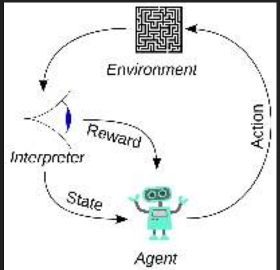
- Supervised learning
  => Statistical regression



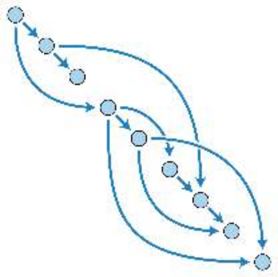
- Unsupervised learning
  => Clustering



- Reinforcement learning
  => exploration / exploitation

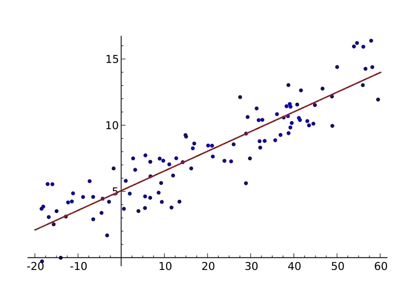


- Symbolic reasoning
  => Graph search

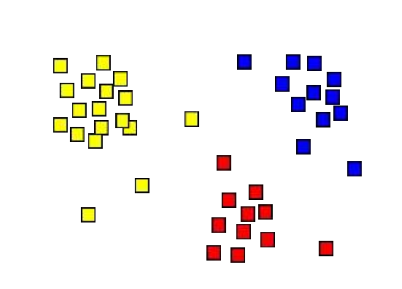


What is AI -- really?

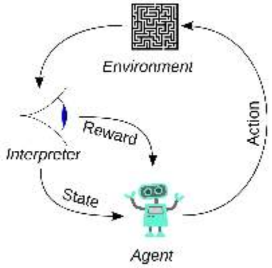
- Supervised learning
  => Statistical regression



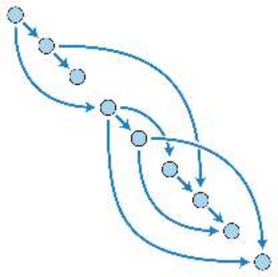
- Unsupervised learning
  => Clustering



- Reinforcement learning
  => exploration / exploitation



- Symbolic reasoning
  => Graph search



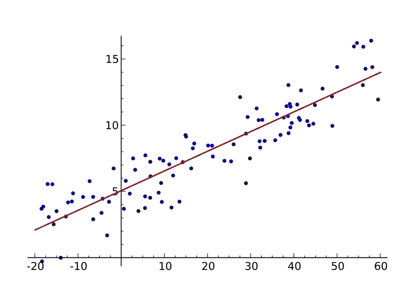
- Bayesian inference
  => basic statistics

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

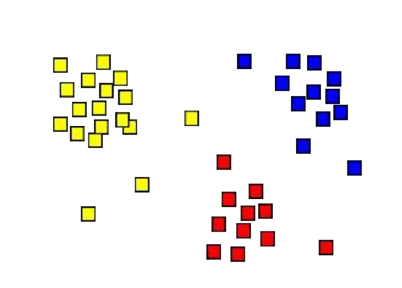
Images: Wikipedia

What is AI -- really?

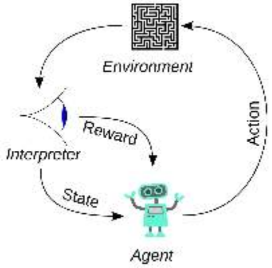
- Supervised learning
  => Statistical regression



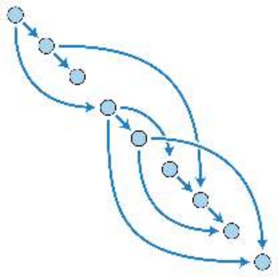
- Unsupervised learning
  => Clustering



- Reinforcement learning
  => exploration / exploitation



- Symbolic reasoning
  => Graph search



- Bayesian inference
  => basic statistics

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

All of these are **necessary** **but not sufficient** for intelligence.

What is intelligence?

How would you define it?

What is intelligence?

https://en.wikipedia.org/wiki/Intelligence#Definitions

- Alfred Binet: Judgment, otherwise called "good sense," "practical sense," "initiative," the faculty of adapting one's self to circumstances ... auto-critique.
- David Wechsler: The aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment.
- Lloyd Humphreys: "...the resultant of the process of acquiring, storing in memory, retrieving, combining, comparing, and using in new contexts information and conceptual skills."
- Cyril Burt: Innate general cognitive ability.
- Howard Gardner: To my mind, a human intellectual competence must entail a set of skills of problem solving — enabling the individual to resolve genuine problems or difficulties that he or she encounters and, when appropriate, to create an effective product — and must also entail the potential for finding or creating problems — and thereby laying the groundwork for the acquisition of new knowledge.

What is intelligence?

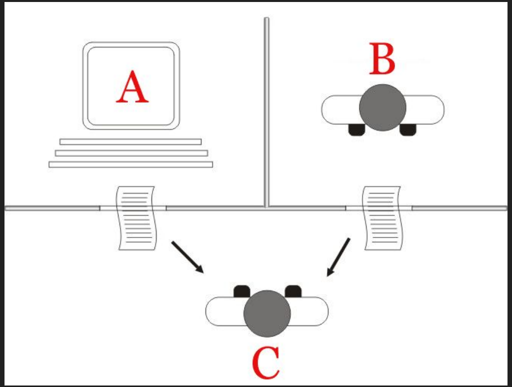
https://en.wikipedia.org/wiki/Intelligence#Definitions

- Alfred Binet: Judgment, otherwise called "good sense," "practical sense," "initiative," the faculty of adapting one's self to circumstances ... auto-critique.
- David Wechsler: The aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment.
- Lloyd Humphreys: "...the resultant of the process of acquiring, storing in memory, retrieving, combining, comparing, and using in new contexts information and conceptual skills."
- Cyril Burt: Innate general cognitive ability.
- Howard Gardner: To my mind, a human intellectual competence must entail a set of skills of problem solving — enabling the individual to resolve genuine problems or difficulties that he or she encounters and, when appropriate, to create an effective product — and must also entail the potential for finding or creating problems — and thereby laying the groundwork for the acquisition of new knowledge.

These definitions (and all others) are **descriptive but not prescriptive**.

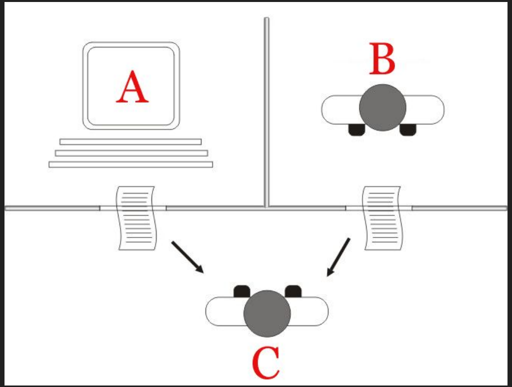
How will we know when we have built AGI?

The Turing test: C has to determine if A or B is a human



How will we know when we have built AGI?

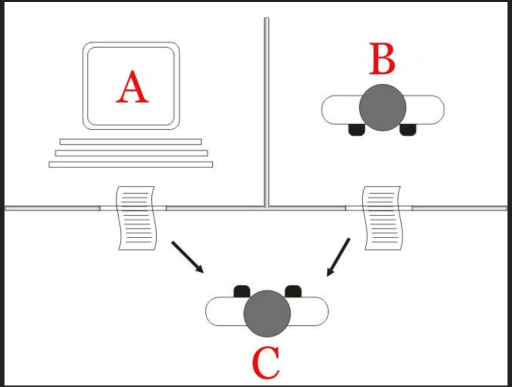
The Turing test: C has to determine if A or B is a human



This only tests the deceivability of the observer (C), not the intelligence of the observed (A).

How will we know when we have built AGI?

The Turing test: C has to determine if A or B is a human



This only tests the deceivability of the observer (C), not the intelligence of the observed (A).

For a superintelligent machine A to pass the test, it would have to pretend to be stupider than it is (“Artificial stupidity”).

What tasks require actual intelligence?

(We are really bad at answering this question.)

FOR HUMANS

		Easy	Hard
FOR MACHINES	Easy		Arithmetic (ever since first computer)
	Hard	**“AI”**	Future AI?

Problems that require actual intelligence to solve need:

- Pragmatic problem solving
- Judicious decision making (=> will?)
- Theory of mind (awareness of others’ minds)
- Awareness of self
- ….

Deep learning doesn’t do any of these things, it is simply statistical regression

And yet we give responsibility for solving problems like this, such as **driving**, to deep learning algorithms to solve.

The result of mis-judging which problems require intelligence

Walter Huang, killed 2018-03-21 by his Tesla in autopilot mode.

Priority inversion

Tesla's statement:

"The driver had received several visual and one audible hands-on warning earlier in the drive and the driver's hands were not detected on the wheel for six seconds prior to the collision. The driver had about five seconds and 150 meters of unobstructed view of the concrete divider with the crushed crash attenuator, but the vehicle logs show that no action was taken."

...Is Elon implying that the Tesla *didn't* have five seconds and 150 meters of unobstructed view of the concrete divider??

The result of mis-judging which problems require intelligence

Tesla's statement, ctd.:

"Over a year ago, our first iteration of Autopilot was found by the U.S. government to reduce crash rates by as much as 40%. Internal data confirms that recent updates to Autopilot have improved system reliability. In the US, there is one automotive fatality every 86 million miles across all vehicles from all manufacturers. For Tesla, there is one fatality, including known pedestrian fatalities, every 320 million miles in vehicles equipped with Autopilot hardware. If you are driving a Tesla equipped with Autopilot hardware, you are 3.7 times less likely to be involved in a fatal accident."

This is a horrendous misrepresentation of the statistics: (1) "equipped with Autopilot hardware" has no relationship to number of miles driven with Autopilot engaged; (2) owners of vehicles "from all manufacturers" is very different than the Tesla owner demographics.

The real stats [2016]: One disconnect every 3.7 miles on average; 3.7T miles/year driven in the US => If all vehicles were Teslas, there would be 1 TRILLION disconnects per year.

But we’re not even properly solving the problems that don’t require intelligence yet

MULTIPLE times a Tesla has crashed into a parked fire truck while in autopilot mode.
What could be more visible to cameras, or radar, *or both*, than a parked fire truck?
Why would a Tesla crash into a fire truck, if it clearly “saw” it?
Why is Tesla overselling this technology, minimizing the seriousness of safety issues, and lying about the statistics?

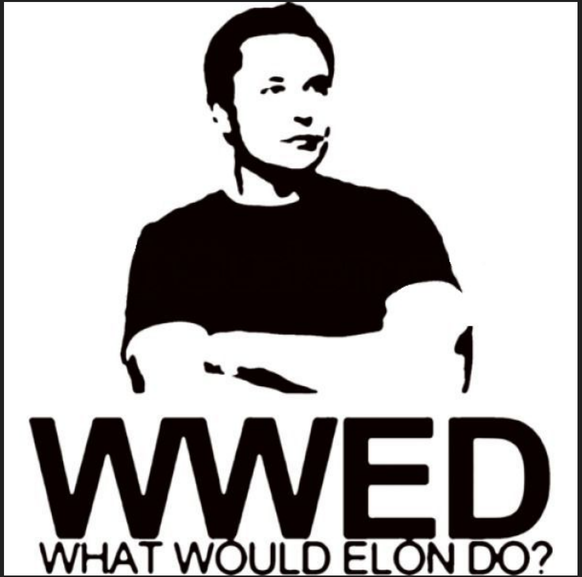
If Elon Musk wants to talk about the ethics of AI...

...he should start by accepting responsibility when AI-powered systems he ships ***kill*** his customers ***for completely preventable reasons***.

We need to start talking about <u>corporate responsibility</u>
in the ethical use of AI.

I took CS classes from four universities, including MIT, and never had to take a single class about engineering responsibility or ethics.

Some of you probably have this on your wall

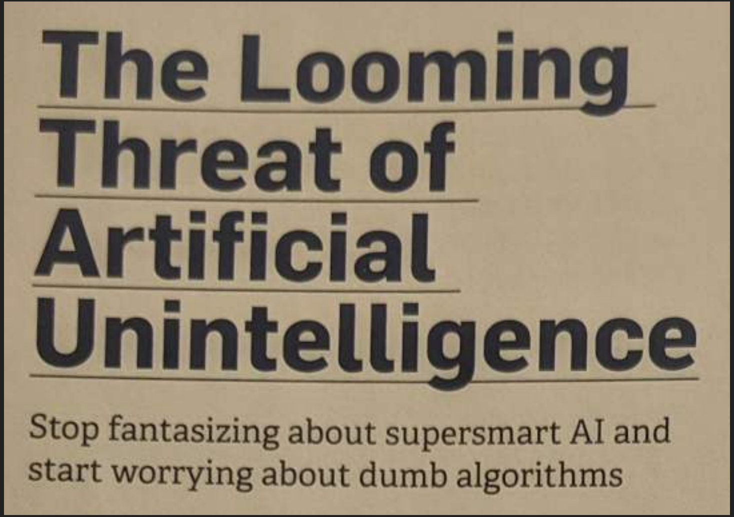


But don’t be like Elon...

Take responsibility for the technology you create

The real dangers of AI

- **Not evil AI, but bad AI:** “...there is a real AI threat, but it's not human-like machine intelligence gone amok. Quite the opposite: the danger is instead [bad] AI. Incompetent, bumbling machines.” [-Motherboard]



The Looming Threat of Artificial Unintelligence

Stop fantasizing about supersmart AI and start worrying about dumb algorithms

The real dangers of AI

- **Bias and discrimination:** People being denied car insurance because a machine learning algorithm noticed that a few very expensive accidents had license plate numbers ending in the digit ‘1’, and decided that this was a good discriminator
    - GIGO: Garbage In, Garbage Out
    - BIBO: Bias In, Bias Out
- **Priority Inversion:** Teslas deciding that lanekeeping is more important than avoiding solid obstacles
- **Placing too much trust in automation through ML**, without appropriate failover and fallback systems (e.g. stockmarket flash crashes)

The real dangers of AI

- **Deepfakes and the coming crisis of reality**
  - Real claims about fake news vs. fake claims about real news
  - What will it mean when you don't know what to trust or believe anymore?
  - A crisis of reality will become a crisis of trust, which will fundamentally undermine society

The real dangers of AI

- **Violation of privacy:**
  - Even if you have nothing to hide, you should still have an expectation of privacy, it is fundamental to living a free life
  - ML and big data analytics are critical enablers of surveillance states
  - How do we balance privacy against security? China’s approach vs. the West
- **Reinforcement of filter bubbles:**
  - People *like* filter bubbles, and ML enables people to reinforce their cognitive biases
- **Targeting of voters and the undermining of democracy**
  - ML-powered disinformation campaigns and election meddling

All of these issues are about the misuse of technology by humans, not the evil intent of machines.

How do we productively think about AI?

AI (ML) is a power tool for the human brain.

Think about **Intelligence Amplification / Intelligence Augmentation (IA)**, not AI.

(1): ML helps us achieve purpose in life.

Look for problems that fit this pattern:

- What do people care about? What do they want?
- How could meeting that need be automated or optimized with ML?
- How can tedious or repetitive tasks be automated?

How do we productively think about AI?

AI (ML) is a power tool for the human brain.

Think about **Intelligence Amplification / Intelligence Augmentation (IA)**, not AI.

(2): AI (ML) allows us to sift through enormous quantities of information quickly.

Look for this pattern:

- What data sources do you have in your organization that you are not currently even capturing?
- Have you ever performed any large-scale visualization of the dataset, to identify correlations between variables? If not, hire a data science team
- You have to talk about data science before you can talk about AI.

How do we productively think about AI?

ML helps us automate some *“snap judgment”* tasks -- primarily:

- Speech processing
- Text processing
- Computer vision (object recognition, OCR)
- Prediction

ML doesn’t work as well for *complex multi-step reasoning* tasks that require *deep understanding of context* or *pragmatic problem solving*.

How do we productively think about AI?

Educate yourself so that you can smooth over the peaks and troughs in the hype, and don’t drink the Kool-Aid (don’t buy into empty hype)

=> You will outlast fluctuations in public opinion about AI (boom/bust cycles, or AI springs and AI winters) better than anyone else.

How do we productively think about AI?

Take responsibility for the technologies your create

Make them safe

Make them effective

Make them fail gracefully without hurting anyone

Don’t oversell or overpromise

And remember:

The sum of

a human brain + a machine

will always be greater than either alone