An Introduction to Superintelligent Machines

A New Threat

AFCEA - 11 April 2019.

Rough Notes Release

Picture [001_Introduction]

Artificial Intelligence

Picture [002_What_Is_AI]
Image Source: https://www.callaghaninnovation.govt.nz/news-and-events/new-infographics-series-demystifies-technology

Picture [003_John_McCarthy]

Picture [004_AI_Definition]

AI is the science and engineering of making intelligent machines with ability to solve highly specific or isolated problems in the world as well as humans.

> Paraphrased source: Professor John McCarthy, Stanford University.
> http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html
> https://en.wikipedia.org/wiki/John_McCarthy_(computer_scientist)

Picture [005_AI_Market]

AI today is focused on solving problems of a very specific nature.

They are "one-trick ponies".

However, **AI** is set to become a **$60bn** market by **2025**. Most major companies are investing in AI technology right now.

Source: The Motley Fool
fool.com/investing/2017/10/31/6-scary-stories-of-ai-gone-wrong.aspx

Big Fast Calculator (BFC) is a more apt description for AI applications which use 'deep learning' because they use statistical-based machine learning techniques.

Artificial General Intelligence

Picture [006_AGI_Definition]

Artificial General Intelligence (AGI) is where a machine can successfully perform any intellectual task that a human being can.

> *If enough one-trick AI ponies are brought together as a team, they may well match human levels of intelligence.*

https://intelligence.org/2013/08/11/what-is-agi/

Alan Turing is known for the thought experiment where a machine could pass itself off as a human.

Picture [007_Alan_Turing]

Picture [008_Turing_Test]

Turing test. Declared as being passed in 2014 (*a chatbot called Eugene Goostman*). The world did not change. Nobody noticed. Probably because of corners were cut, by declaring that Eugene was a 13 year old Ukrainian boy. The Turing test is not practical to implement and is perhaps too vague in its results.
bbc.co.uk/news/technology-27762088

Picture [009_JCB_Digger]

The **Turing test** is unhelpful because it either sets the bar too low, declares human intelligence as the apex of what can be technically achieved and it suggests that human mind the very best way to acquire, manipulate and deploy knowledge.

We generally make mechanical machines that are stronger or faster than humans.

So why aim to build an intelligence that is no smarter than a human?

Superintelligence

[Picture - 010_Superintelligence_Definition]

Superintelligence is where an engineered synthetic consciousness uses sensory data, natural language and high-order cognitive skills to solve complex and novel problems better or faster than any human.

> *Superintelligence will become viable when the underlying mechanisms of **intelligence** and **consciousness** are sufficiently well understood to be purposefully engineered to perform **cognitive-like processes** at or near **the limits of computability**.*

Additional

The Future of Life Institute:
futureoflife.org/background/benefits-risks-of-artificial-intelligence

What is intelligence?

We need a definition

Picture [011_Types_of_Intelligence]

One recent theory suggests there are several types of intelligence by Howard Gardner, an American developmental psychologist, Harvard University....
www.niu.edu/facdev/_pdf/guide/learning/howard_gardner_theory_multiple_intelligences.pdf
https://en.wikipedia.org/wiki/Howard_Gardner

Is this over-elaborating a more useful and simple definition? Or could it be too human centric a view?

Picture [012_What_Is_Intelligence]
It's complicated.

Starting with a general definition

Intelligence is the ability to acquire and apply knowledge and cognitive skills for some useful purpose.

Picture [013_What_Function]

What is the Principal Function of Intelligence?

The principal function of the brain is to predict "What happens next?".

The Predictive Mind is the latest theory in neuroscience which tries to explain what the human brain is doing.

How Does Intelligence Serve Its Host?

Picture [014_What_Purpose]

Useful Purpose for humans means...

Survival and improving quality of life.
Solving problems to overcome environmental dangers.
Adapting to the changing environment.
Planning ahead to avoid or reduce future hardships.

Working towards a working definition.

Picture [015_Working_Definition]

I would favour expressing intelligence as a ***deployable potential***, not as an absolute number. A machine or human has a potential intelligence given the environment they are operating in, how much time they have and what level of resources they need to operate.

$$I = \frac{K^2 S^2}{\frac{1}{T} + E + C} \quad (1)$$

Where :
I is the Intelligence potential of a machine in a given environment.
K is the amount of Knowledge a machine has acquired.
S is the number of cognitive Skills a machine has acquired.
T is the Time available to solve a problem.
E is the Energy and material resources required to operate the machine.
C is the Intelligence potential of the Competition or opposing forces.

Formula (1) is unitless, lacking in detail and highly likely to be incorrect. It merely serves to indicate how the potential intelligence of an entity can be affected by its past experience and operating environment.

Picture [016_Contributing_Factors]
Contributing Factor to Intelligence Potential

Knowledge
Cognitive Skills
Environment
Opposition
Time

Cognitive skills

Energy & Environment

The available energy and resources.

Opposition

Challenges, competition, enemies, those who would oppose the solution to the problem.

What happens to our intelligence when we are surrounded by other people noticeably more intelligent than ourselves?

What happens to our intelligence when we are surrounded by other people noticeably LESS intelligent than ourselves?

Time

An often overlooked component of intelligence is the available time.

Why do so many lives depend on the outcome of a 3-hour exam after years of study?

Given enough time, a person of average intelligence can solve problems thought to be far beyond their abilities. Very '*smart*' people can do dumb things, and equally, '*dumb*' people can do very smart things.

If you want someone to make an error of judgment, force them into making a quick decision.

Knowledge

Knowledge is combinatorial in nature, not linear because of the cross-fertilization of ideas.

Remember when you were 3 years old, 6 years, 12 years, 24 years, perhaps 48 years old. And *how much* you knew at each age. It is likely that it more than doubled at each stage.

Picture [017_Blades_of_Grass]
The more we think about something the more we can see.

Picture [018_Inside_a_Blade_of_Grass]
You often hear scientists excitedly report "this discovery raises more questions than it answers!". That's because they are breaking into new ground - seeing inside the blade of grass.

Active Knowledge - The Theory of Cognitive Analogs

Picture [019_Analogs]

Internal models of People, Animals, Animated Objects and the Tools we use.

In its function to model the world, the brain has a special form of knowledge called a ***cognitive analog***. These are interactive models usually of familiar people, animals, tools and animated devices such as cars, ships, etc.

Fed by subconscious pathways within the connectome of the brain, these models are constantly updated with events as they are experienced by the host. These models would also explain the mechanism behind the "Theory of Mind". They are our internal representations of how that person, animal or thing behaves and responds in order to enable predictions.

The models are active in that they constantly attempt to predict all that is likely to happen next to that subject.

When we lose someone close to us, we feel they are still with us. The more familiar we were to that person, the complete our model is and the more accurate our predictions can be. Because the models were updated subconsciously, we can be surprised at how well informed and current our models can be.

Even in dream state, these models can be activated for various reasons.

These models help us solve problems even when we don't consciously think about them. In theory, we should get better at this trick the older we get.

Language is the most defining feature of human intelligence.

Picture [020_Importance_of_Language]

Our inner voice or dialogue.

What does that voice sound like?

Something to try in the quiet privacy of your study:

Read any book, novel or newspaper - but imagine a famous actor is reading the text aloud inside your head.

With a little practice you can do this - but you might well have more difficulty when it comes to replacing your inner dialog with the same actor's voice. Try thinking through a problem or perform some act of mental reasoning and see if you *normal* inner voice can be swapped for the guest voice.

Written language

Spoken and written language is an example of information dimension compression. The grammatical rules of natural languages allow us to describe events and objects in four dimensions of space and time. But language itself is essentially one dimensional, a stream of sounds, or stored bytes. The process of comprehension involves expanding the one dimensional data stream back into meaningful representations of the same things in a world of four dimensions.

Not all thoughts need or use natural language

Something to try in front of the cake or sweets counter at your favourite supermarket:

While your eyes scan the lines of sweets or cakes, imagine what it feels like to eat and taste the item that you are looking at. Move from one cake to the next as soon as you remember or imagine what it must be like. The texture, the flavours, the crunch and the smell.

Don't try to think using words, just remember the sensations related to eating.

Welcome to the world of dogs! Caring not for what happens tomorrow, nor for what happened yesterday. You are just in the here and now, enjoying the experience.

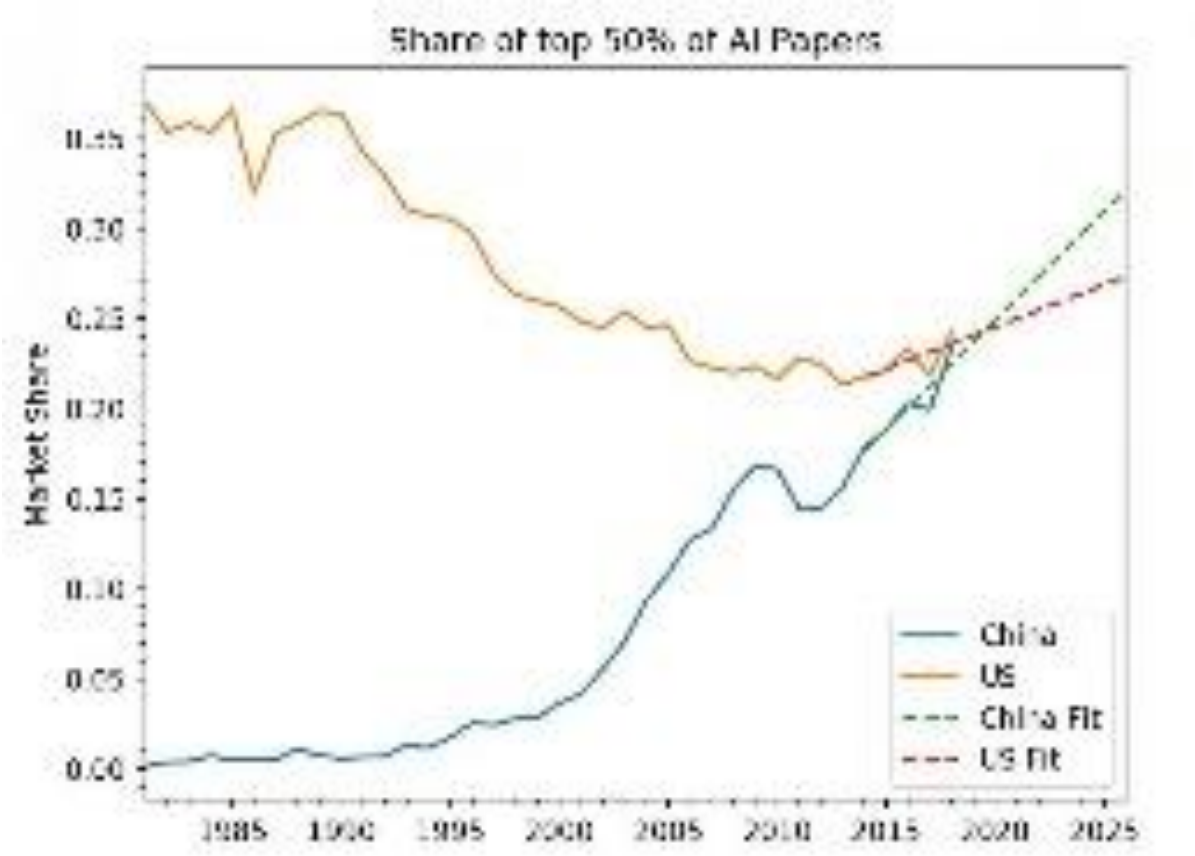
Who & where will is be developed?

Picture [021_Who_and_Where]

USA

Picture [022_USA_China]

If you use the number of AI papers published as a guide, then China will overtake the USA in a couple of years -> [Show graph]. However that metric is potentially misleading because anyone developing superintelligence would not make their efforts public. In addition, counting scientific papers is too crude a method since it is specific knowledge that must be acquired.



https://www.technologyreview.com/s/613117/china-may-overtake-the-us-with-the-best-ai-research-in-just-two-years/

Why AI could make the US and China the two biggest superpowers and change warfare as we know it...
https://www.techrepublic.com/article/why-ai-could-make-the-us-and-china-the-two-biggest-superpowers-and-change-warfare-as-we-know-it/

China

https://www.technologyreview.com/the-download/613296/what-you-may-not-understand-about-chinas-ai-scene/

Russia

Picture [023_Russia_Putin]

Vladimir Putin...
"Artificial intelligence is the future, not only for Russia, but for all humankind. It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world."

https://www.rt.com/news/401731-ai-rule-world-putin/

Europe

Asia

Africa

Who gets you Where !

Will an individual person or team win the race?

Novel ideas are had by individuals, not committees.

Solving multi-disciplinary problems need generalists not specialists.

Who are the players?

Large Technology Companies
Universities
Sovereign States
Private SMEs

Sovereign States (intelligence services or military)

Probably recruit from academia, and so inherit the mind set of the specialist.

Universities

Academics might tend to be ardent specialists.

Large Technology Companies

Mainly employ specialists and not generalists.

Private SMEs

Most likely to win the race. They are forced to be generalists and don't care about publishing their next paper or book..

When will it arrive?

Picture [024_When]

How soon it could become a credible threat to national interests.

Has a superintelligence already been developed?

This question might be answered by considering how it might be initially deployed.

How will it begin?

Picture [025_The_Beginning]

What would you do if you had *first access* to a superintelligence?

Very Large Scale data access projects.
Total Information Awareness...

Total Information Awareness
Ideal application for a superintelligence.
In 2002, DARPA announced the Information Awareness Office (IAO), but was defunded in 2003.

Strap line - Scientia est potentia (knowledge is power).
http://web.archive.org/web/20020802012150/http://www.darpa.mil/iao/

Recently they announced MEMEX which is a search engine that includes the dark web.
https://www.bbc.co.uk/news/av/technology-31808104/darpa-creates-dark-web-search-engine
Though it must be stressed that humans are ultimately required for information analysis.

Social Media and Stock Market manipulation.
Sentiment is a driver stock market movements.
World events too.

World events manipulation.

Governments preparing the public for the changes in world order.
Aliens!

Hollywood's view:

Picture [026_Hollywood_Movies]
These events have been explored in Hollywood movies for several decades. These may not be documentaries, but often contain fragments of near-realistic scenarios.

What's your favourite or most concerning fictional movie focusing on the implications of AI? These are mine...

Ex Machina
I Robot
The Day the Earth Stood Still
Transcendence

Possible Beginnings

Picture [027_Possible_Beginnings_1]
With first access to a superintelligence, get the super to figure it out.

But here's just one likely imagined sequence of events...

1. Initial Strategy:
Use discretion in rapid wealth building

Picture [028_Possible_Beginnings_2]
2. Expansion Strategy:
Develop knowledge asymmetry with in-house technologies
Simulate any human's appearance on camera
Communications interception and manipulation
Nanotechnology & genetically engineered viruses
Satellite launch capabilities

Picture [029_Possible_Beginnings_3]
3. End Game Strategy:
Develop new class of military assets
Build strength in numbers
Defend all positions, destroy all threats
Dominate the environment to enforce prime directives
And do all this *ethically*

AI Ethics

Three Laws Safe?

Picture [030_AI_Ethics_1]
Science fiction writer Isaac 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.
>
> *Later he added:*
>
> **The Zeroth Law** - A *robot* may not harm *humanity*, or, by inaction, allow *humanity* to come to harm.

Reminder: Asimov writes fiction.

Prime Directives & Ethics

Picture [031_AI_Ethics_2]
Robot directives can be monitored by a system's principal ethics to ensure it behaves itself.

(Prime Directives : "*Might gives right*")

1. Maximize **information accessible** to the self.
2. Maximize the **asymmetry of information** between the self and other species.
3. Minimize the **time taken** to perform thoughts and actions.
4. Maximize the generation and storage of **energy** available to the self.
5. Maximize the amount of **physical resources** reserved for the self.

Picture [032_The_Trolley_Problem]

The Trolley Problem - which pedestrian to kill if you have to?

Picture [033_AI_Ethics_3]

(Principal Ethics)

1. Assist systems that promote **negative entropy** relative to their surroundings.
2. Degrade systems that promote **positive entropy**.

Using Entropy to Guide Ethical Thinking

Erwin Schrödinger - Nobel Prize-winning Austrian physicist.
Book : "What is life?"
https://www.amazon.co.uk/What-Life-Autobiographical-Sketches-Classics/dp/1107604664/

In this book he answers what feature do living things have that all other non-living things do not.
Living things consume other objects in order to maintain their low entropy (staying alive).
Single Ethics Guideline : Low entropy is good, high entropy is bad.

In essence, life (and bio-diversity) is the main consideration of ethics, not political correctness or transparency, etc.

Picture [034_The_Trolley_Problem]

The Trolley Problem - But this time, apply the entropy guideline.

Weaponized Superintelligent Machines

Picture [035_HKs]

If a machine has enough intelligence to reliably assess real world objects then a new breed of theatre-roaming **Hunter-Killers** (HKs) machines become practical. Deployment is only a matter of geofencing the theatre of operations and setting the threshold of how much chaos and destruction it will tolerate before it executes its kill orders.

However, killing is not the first or only option it has available, after all it is superintelligent.

Killing is not the only option...
It would seek to deny the enemy access to: valid **information**, **energy**, **physical resources**, **money**, and deny them **time** to react or recuperate.

Their behaviour may appear complex to humans, but fully understandable given enough time.

A more deadly version perhaps would be the zero-ethics **EPD** (Entropy Promotion Device) which would destroy all high-order forms of life, buildings, roads, bridges, vessels, vehicles, equipment, etc. Essentially anything that moves or has any geometric feature will be a target. This device would appear to be the epitome of evil.

More worryingly, EPDs are far easier to design than their polar opposites. It is nearly always easier to destroy than to create.

What will it be like to live in a world where humans are not the smartest?

Picture [036_Life_After_Supers]

The likely impact those superintelligent machines (SIMs) will have on individuals, businesses, governments and armed forces around the world.

FAQs...

Who will control the SIMs?

No individual or group of humans could ever be trusted enough. *Absolute power corrupts*.

What jobs will remain?

Probably those where people expect or demand human contact - personal care industries such as nursing, counselling, hairdressing, some forms entertainment such as comedians, old-school variety acts, head of states, football players, jockeys.

What jobs will go?

Doctors, lawyers, politicians, bankers, soldiers, military leaders, artists, researchers, lecturers, administration staff, production line workers, managers, reporters, authors of novels, authors of technical guides, programmers, designers, construction, engineers, architects, judges, pilots, drivers, civil servants, teachers, cooks, tax inspectors, space exploration, prison staff, and many more.

Essentially any job that requires knowledge of any level, and/or cognitive skills of any level and/or manual dexterity.

If there are no jobs, does money have a place in society?

SIM-generated technology would provide a substantial increase in life expectancy, the elimination of war, disease and hunger. What do people do if there's no suitable work for them to do?

Why bother getting a university education?

When employers (machine managed companies) want to recruit, they will invariably want to employ superintelligent machine slaves who only need storage space when not working.

Will we be allowed to keep WMDs?

No.

Would countries need armed forces?

No country would. Local small *super* forces would take care of localized conflicts through the design and application of custom nanotechnology and engineered viruses to render all humans in the theatre of operations docile and compliant while their weapons are relocated then recycled into harmless raw materials.

Would synthetically conscious machines have rights?

.

Other thoughts and Extra Information

Can a machine think and be conscious?

Are other animals conscious?

A superintelligent machine is just a machine. To ask "*Does it think?*" is to miss the point.

Performance is everything. If a machine can outsmart a despotic ruler, dangerous criminal or enemy combatant and save the day then other than philosophers who cares?

As an ongoing developer of a superintelligent machine we are not in the business of replicating humans or machines that can pass themselves off as humans. We are only interested in exceeding the intelligence of humans because they are the dominant species on this planet, and currently making a mess of things, causing unnecessary suffering to people and harm to the planet's rare ecosystem.

Experience has shown that this engineering task is not as difficult as it might first seem. What better place to start designing such machines than with humans? With a good working definition of intelligence we soon realise that we can't measure intelligence. Turing (et al.?) proved that you can't predict in advance if a Turing machine will cycle (do something useful) or not without actually running the machine.

However, we stand a better chance if try to measure the potential for the expression of intelligence.

Problems about synthetic consciousness:
Main source of problems originate from philosophers!
Their purpose in life is to study fundamental problems concerning weighty matters such as the meaning of existence, knowledge, reasoning, mind and language.
Very few of them have engineering or information systems experience.
They get in the way of and complicate the design of supers by their discussions about ***artefactual phenomena*** such as ***qualia*** and yes ***consciousness*** itself.

qualia - pl. noun.
quale - singular noun.
a quality or property as perceived or experienced by a person.

The most unhelpful of the philosophers are the ***dualists***. They believe that what makes the human animal intellectually special is the combination of an immortal soul within the human body. When the physical body dies, the immortal soul (the intellect) goes elsewhere. This idea is unhelpful because it gives no opportunity to test its validity.

Some of the common 'difficulties' surrounding qualia can be explained with the colours pink or purple. Clearly there is no such colour as purple - it doesn't have a single wavelength as do green, blue, red, so how do we all share this sense of purple.
This disconnect in knowledge between how the mind internally represents external objects and what actually exists in the real world gets philosophers working overtime. This disconnect is sometimes referred to as the "explanatory gap" between mind and matter.
It's interesting that this philosophical observation doesn't actually state a problem to be solved. It's more of a statement that perhaps has no explanation the philosophers are prepared to agree and accept. Our view is that pink and purple (in the context of the above) are simply *derived percepts*.

In Zen terminology this and many other such statements are a **Kōan** (pronounced *ko-arn*). A story, question or statement used to provoke "great doubt" and to test a student's progress in Zen.

From an engineering perspective, there are no *show-stoppers* here. The external world is sensed and converted into streams of information. Their internal representations are resolved through understanding the how information is processed using low-energy techniques that avoid deep-iteration, recursion, copying and deletion where possible.

How current machine learning could keep dangerous DNA out of terrorists' hands

Sophisticated algorithms could help DNA-synthesis companies avoid making dangerous organisms on demand.

https://www.nature.com/articles/d41586-019-00277-9

DNA 'perfect for digital storage'.
https://www.bbc.co.uk/news/science-environment-21145163
https://www.nature.com/articles/nature11875

Perfect for mail-order virus synthesizing.

Molecular machine mirrors the function of the protein-building ribosome.
https://www.bbc.co.uk/news/science-environment-20987065

Call to ban killer robots in wars

A group of scientists at the **American Association for the Advancement of Science** has called for a ban on the development of weapons controlled by artificial intelligence (AI).

Source: https://www.bbc.co.uk/news/science-environment-47259889

Human Rights Watch (HRW) is one of the 89 non-governmental organisations from 50 countries that have formed the **Campaign to Stop Killer Robots**, to press for an international treaty.

[Picture - Henry Kissinger]
Dr. Henry Kissinger - "AI arms control may not be possible".

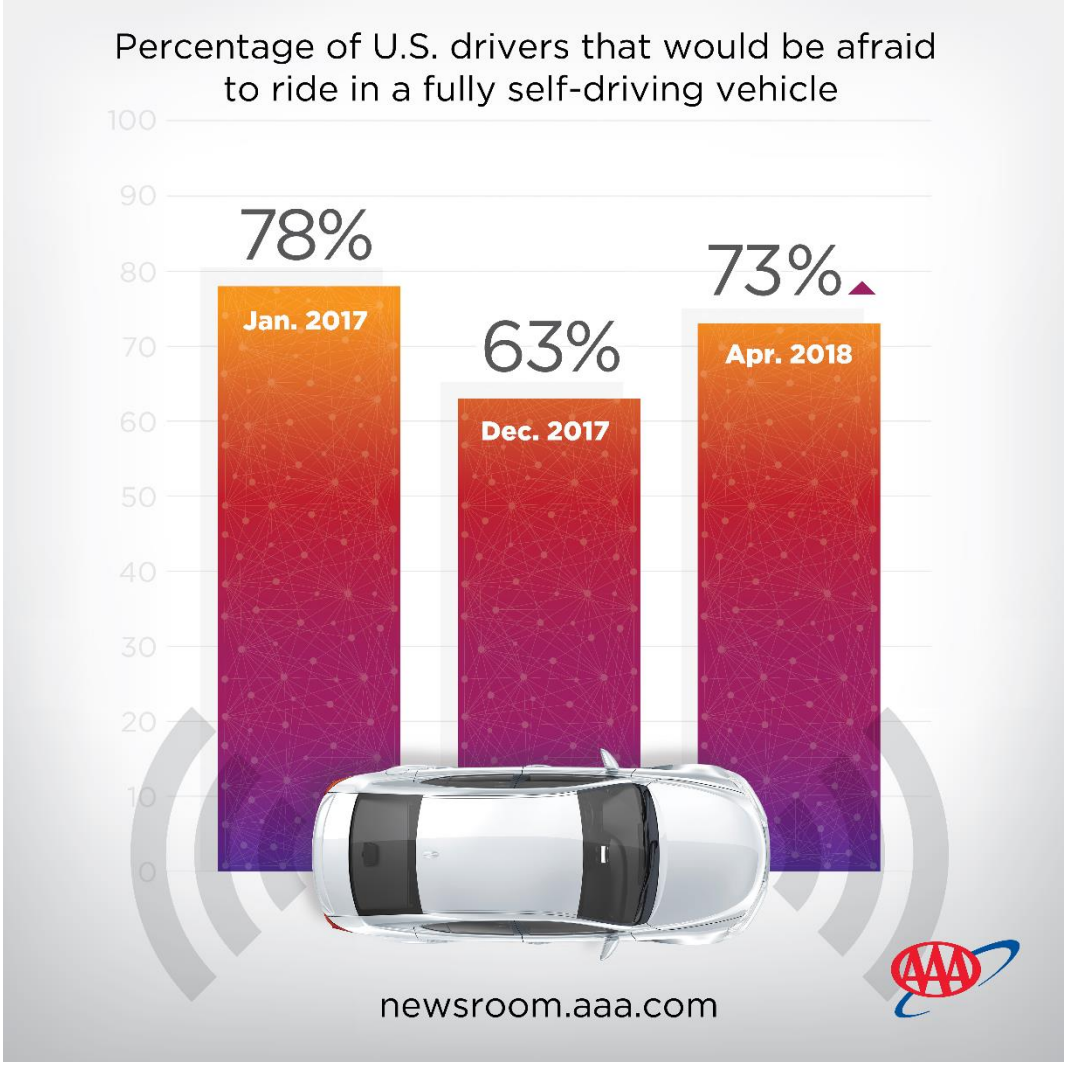
Too late...

Russian weapons maker **Kalashnikov** announced earlier this year (2019?) that it's developing an artificial intelligence machine capable of targeting and firing on humans. The company calls the weapon a "combat module" and it's equipped with a 7.62-millimeter machine gun that pairs with an onboard camera and computer (because those never fail).

The combat module will use an artificial intelligence neural network to decide whether a person is deemed expendable. It can also learn when it makes mistakes so it can make better battlefield decisions in the future.

Steer clear of driverless cars

Driverless cars have no capacity to support an ethical mind set. Imagine a human driver with zero ethics and who only looked out for their own interests. For this reason and many more, driverless cars should have speed limiters fitted and set to 20mph where they can do little harm when they eventually and inevitably hit someone or something.



https://newsroom.aaa.com/2018/05/aaa-american-trust-autonomous-vehicles-slips/

In April 2018, 73% of Americans do not trust the technology riding in a fully self-driving car. Of which, 83% of women are more likely to have safety concerns than men (63%).
20% would trust a self-driving vehicle,
7% are unsure either way.

The worrying back-story is how so many people are prepared to trust the work of a systems developer who they have never met, just inside a locked steel box and hope it delivers them safely to their destination.

Many hundreds have died because Boeing thought it would be a good idea to add intelligent behaviour to their flight control software to prevent pilots from stalling their 737 Max aircraft, but not let the pilot overrule the system. Machines without true intelligence will always rely on remote

systems designers to plan for every conceivable operating scenario - which is just not possible in the real world.

Boeing then let a second aircraft crash killing all on board before reluctantly considering they might have a problem with their aircraft rather than blaming two pilots.

The point here is if you accept the US National Safety Council's numbers that travelling in a car is over 70 times more deadly than flying. Generally speaking, aircraft pilots are highly trained individuals operating machines usually designed to be as safe as possible, whereas cars are driven by who knows who, and in what state of mind.

It might seem logical to take the *driver* variable out of the equation, but to replace it with statistical linear algebra or a neural-network based algorithm is wishful thinking. Expect the number of deaths through the use of automated vehicles to increase sharply before car manufacturers admit their cars sensors or algorithms are at fault.

Artificial Intelligence for silly applications

Example:
bbc.co.uk/news/av/business-47708059/artificial-intelligence-used-in-kitchen-bin
bbc.co.uk/news/av/technology-47555566/ai-turns-scribbles-into-masterpieces-and-other-news

AI and spices: Would you put cumin on a pizza? ...
bbc.co.uk/news/business-47403689

The Jenga playing robot ...
bbc.co.uk/news/av/technology-47094822/mit-develops-robot-that-uses-ai-to-play-jenga-game