The Hidden Dangers of AI

Have the very first shots already been fired in the robot revolution? A **recent report** by the UN says that an autonomous quadcopter drone built by a Turkish company "hunted down and remotely engaged" retreating rebels without being directed to do so, after a firefight in Libya in March, 2020.

It's not known how effective the attack was, but the **Kargu-2 drone** can function manually or on its own, day and night, with a variety of weapons, and even blow itself up in a kamikaze attack. The immediate question (as with all drone attacks) is how well it could distinguish the bad guys from the innocents.

How could it be sure? What triggered it to shoot? Just the idea of autonomous killing machines has troubled many leaders. The UN and EU Parliament have opposed killbots for years, and a **worldwide campaign** to outlaw killer robots was signed by 4,500 AI experts. Stephen Hawking, Elon Musk and many others have also strongly **urged caution**.

Could ***The Terminator*** come true? That movie and its franchise was far from being the first to predict a bloody revolution by our mechanical servants. Robot rebellion is a very old theme in sci-fi. In fact, the obsolescence of human beings and such a revolt was originally first foreseen an entire century ago, in ***R.U.R.***, the play that invented the term "robots".

The scary thing is that the more we develop **AI** (artificial intelligence), the greater those anxieties become. For there are some very good reasons to distrust AI solidly based in how it actually works.

Making machines think

The dream of thinking machines is very old, and the quest to make them cognitively capable has been long and grueling, with far more failures and dead ends than successes. The holy grail of "**general**" AI which would think as broadly and nimbly as humans can is still quite distant: all the achievements of AI have been very limited and highly specialized.

Yet AI systems can be immensely powerful. Computers are tireless and relentless, never losing attention, able to process far more data than humans can, far quicker and in exquisite detail. Their potential and promise is so great that we humans are constantly throwing new problems, tasks, and tools at them.

We tend to regard AI systems just like any other computers – objective, emotionless, and logical. But we are discovering that intelligence is quite complex and subtle, and it turns out that some of the problem is due to how machines are taught to think.

Why? Because the digital universe is absolute: all 1s and 0s, everything is black or white. Yet the real world is not only full of grays, but glorious colors. Fixing computers to handle all that uncertainty is hard.

The universe is a messy, unruly place, and we only got to where we are by intuition, creativity, and luck. We cannot tell computers how to be intuitive partly because we cannot precisely define it ourselves. So designers had to be creative, even tricky, in teaching machines ways to mimic human activity.

Take walking, for instance. While instinctive in us, it still takes a lot of falling down before we master it. Walking requires motive, perception of obstacles and ways around them, balance and timing, plus moving our limbs and knowing where they are at all times.

Originally it was felt that feeding a robot a program that calculated every possible variable would be enough. But it's impractical to cover everything save in the most constrained situations. So it was realized eventually that the best way to teach machines was to make them able to teach themselves.

That's how humans mostly learn, but even so, AI's digital exactitude is extremely confining. A way around it is called **deep learning**, using probability and testing the results. Instead of applying rigid rules, the machines try to find common **patterns** in data, judging how likely their perceptions are, while feedback loops constantly adjust how they perform.

So the amazing feats robots do in **videos** are not simply programmed into them; like human tricks, they are the result of countless hours of practice.

Likewise, with the computer's ability to quickly process data, we can feed AIs huge amounts for them to sort out on their own. To teach how to distinguish a dog from a cat, for instance, researchers can input thousands of pet photos with only the correct identification given. Without any instructions,

the system then compares images to puzzle out which ones are puppies and which are pussycats.

Then they are tested with unidentified samples to see how well it worked. But the AI likely cannot tell us *why* any more than a toddler could, and it is quite possible that the computer chose some feature the human programmer never thought of or could predict because the machine is a **black box** whose inner functions cannot be directly observed.

The method works well but has had some pretty bizarre results. Perhaps the computer identified dogs by their tongues hanging out, which would give the right answer most times, but not all. Or perhaps an abstract pattern that looks nothing like a dog might somehow trigger it, which has also happened.

This could lead to bad, even fatal mistakes. Some driverless cars, for example, could be fooled by a sticker on a stop sign, not recognizing the danger. But that is merely the proverbial tip of the iceberg.

Building in bias

Humans are emotional creatures driven by need and experience, shaped by inherited beliefs, and judging the world by how we fit in. All our thought patterns and behavior are shaped by these forces, often unconsciously – and therefore, uncontrollably.

However noble the efforts of engineers to eliminate bias in all its forms, it is **surprisingly difficult**. There are an ever-growing number of disturbing examples of human prejudices that machines pick up.

It doesn't take much. One AI program demonstrated **prejudices** against women and black people just by reading random postings on the net. Another study found that similar outcomes happened when a computer simply taught itself to read English.

The AI system that New York City uses to place students in schools was shown to **screen out** black and Latino teens for the top schools in favor of white and Asian students based on a single test score. Likewise a program used to **predict** future risks of committing crimes to help in sentencing was prejudiced.

Amazon **scrapped** a promising résumé-screening tool that showed bias against women for tech positions as most of the job applications it was trained on came mainly from men, who still overwhelmingly dominate most high-tech industries.

At the same time, Facebook's AI chose what **political ads** users saw, biased in favor of conservatives. Many cities have secretly experimented with **predictive police** software from a private company to plan responses for different neighborhoods by tracking petty crimes like graffiti. Who can say how that has impacted widespread police overreactions?

Algorithmic biases, the prejudices AI acts upon, are becoming ever more dangerous as we give AI more and more power to make decisions that affect our lives. Yet biases are often unseen, particularly by people who share them. It is the proverbial situation of "garbage in, garbage out", fixable *only* if those garbage assumptions or data can be recognized.

All biases are grounded in how the world was assumed to work in the past; algorithms need to reflect how we want it to work in the future. But if it's so difficult to remove unconscious preconceptions, what happens when AI is *deliberately* turned to evil?

Tay, for example, was an AI conversation agent, a **chatbot**, that Microsoft released on Twitter in 2016, designed to mimic a teenage girl. It had to shut down, not once but twice, after users taught it to reply with foul, racist, and sexually explicit language for laughs. Its successor, **Zo**, didn't fare much better.

Then there are **deepfakes**, where an AI program replaces one person's likeness or voice with another. Though their potential for causing political mayhem has been terrifying experts for years, they've mainly been used for porn. However, they are getting easier to do and better all the time, and now even **maps** and satellite photos can be deepfaked as well.

Too smart to be controlled?

Even if AI systems could be made safe, just becoming smarter than humans could lead to our extinction. Many thinkers worry that AI could end us out of pure benevolence (as the machine defines it), maybe even before it becomes **superintelligent**.

A system called **GPT-3** can generate **documents** based on a few words that are so well-written they are virtually impossible to distinguish from human compositions – even poetry and computer code. GPT-3 is so smart because it was trained on hundreds of billions of words. Its potential for harm by deepfakes, or just by eliminating writers, is so great that both access and details about it are highly restricted.

It makes this poor writer recall the translation a far greater one did of a mysterious inscription appearing on a wall: "Measured, measured, numbered and weighed." Any doubts might already be too late.