# Undercurrent

*Philosophical Perspectives on Topics of Interest*

*Untitled, Daniel Roizer*

## War Drones will be the Death of Us

October 16, 2016 Long

After years of promises, artificial intelligence (AI) is beginning to impact the world. But it also brings significant risks in its wake.

Last year, a group of thought-leaders of high technology issued a warning that AI posed considerable dangers to mankind, which needed to be addressed. Signatories included such luminaries as Stephen Hawking, Elon Musk, Steve Wozniak, Daniel C. Dennett, Noam Chomsky, and more than twenty thousand others who signed it online. This raised many eyebrows, as most people still see the idea of robots taking over the world as clichéd science-fiction. After all, computers have been part of our lives for a generation and show no signs of mounting an attack on us.

Is this just scaremongering, or should we be worried about what might be coming out of AI labs in the not too distant future?

In fact, current trends in AI point to a clear direction of travel. The journey has been slower than idealists imagined in the 1970s when the British Government's Lighthill Report hammered the optimism of robotics researchers. Slow though it may be, the journey toward autonomous, intelligent computer systems is getting to the stage where we need to wake up and pay attention. An 'autonomous' machine is one that observes its environment, makes decisions about what action should be taken, and then executes that decision without referring to a human authority. Such machines would, of course, be designed and built by humans (or, in the foreseeable future, they might be built by other machines, which in turn are ultimately the fruits of human labour) but, once booted up, they could carry on without referring back to their human creators.

The advent of driverless cars on public roads and pilotless drones in civilian airspace has jolted some people into realising that a lot has been done in AI research below the radar of the newspapers. Already one person has been killed by a Tesla driverless car, although this was an accident, not an attack. The ubiquity of automated call centres is a bane of modern life but we know that somewhere inside the system there must be people in charge… right? Yes, but for how much longer? We have all had the experience of 'the computer says no': when the bank software says you are ineligible for a loan and no amount of pleading with the branch staff can override the computer's decision. In fact, there is no fundamental reason for banks to employ humans at all. Most interactions occur through the internet or automated call centres and branches now promote automated teller machines.

This direction of travel is taking us straight toward a situation where autonomous machines will exercise substantial control over our lives, not just over our money but extending to the power to kill us. In 2012, the group Human Rights Watch considered automated war-making to be urgent enough to issue a fifty-page report against killer robots. The use of combat drones, with increasing degrees of autonomy, has continued unabated. At three successive UN meetings to debate 'killer robots', the British Government has voted against a ban on the development of autonomous war-drones. The US Navy has been at the forefront of developing lethal drones, with autonomous boats and submarines and airborne drone swarms.

As in any slippery slope, the first developments are so sensible they seem inevitable: underwater drones can operate in environments difficult for human troops, and there is little risk of collateral damage to civilians and the associated bad press that surrounds robots autonomously taking human lives. Earlier this year, US

Defense Secretary Ash Carter announced a budget of $600m for submarine combat drones. Drone warfare in the air is also getting big funding. Driving this is the fear of 'capability surprise'—a fear coming from the realisation that US warships are vulnerable to attack by drone swarms, and the only credible defence is a counter-swarm, which is why projects to demonstrate swarm warfare are already commencing. Combat drones are coming to a battlefield near you soon.

The AI community has slowly woken up to the risks of war drones, with idealistic campaigns to outlaw them, such as the Campaign to Stop Killer Robots. In December 2015, the Leverhulme Trust gave $15m to an AI research centre at Cambridge University for research on "the long-term safe and beneficial development of AI" including "near-term challenges such as lethal autonomous weapons". In January this year, Elon Musk donated $10m to the Future of Life Institute "to run a global research program aimed at keeping AI beneficial to humanity." Note the buzz phrase "beneficial AI" as opposed to – what? Destructive AI, presumably? And in July this year, Google Brain in collaboration with Stanford University, University of California, Berkeley, and OpenAI, published a 29-page analysis of AI risks.

Media attention has dwelt on the mere fact that combat drones can autonomously kill human beings, as if this marked a radical departure in the moral and legal framework of war. Slaughter by machines is politically contentious but not, in itself, philosophically problematic. Bombing urban areas will kill many non-combatants, euphemistically referred to as collateral damage. The human bomber does not choose who dies: that is determined by the bomb and the laws of blast mechanics. Whether you die at the hands of military personnel following orders, or a robot following instructions is somewhat academic. What should be worrying you is: whose instructions will the robots be following?

Even the Financial Times, known for its astute recognition of trends, is not keeping up. A few days ago, Robert Shrimsley, Managing Editor of FT.com, found cause for derision rather than sober reflection in last month's formation of the *Partnership on Artificial Intelligence to Benefit People and Society* by Amazon, DeepMind/Google, Facebook, IBM, and Microsoft. Its tenets mention a range of industries in which AI "can be leveraged to help humanity" and it supports "promoting safeguards and technologies that do no harm" but ominously omits any mention of AI's biggest threat to humanity: intelligent weapons of war.

## A Worst-Case Scenario

To a military mind, an intelligent combat drone is just another soldier. You train it, discipline it, make it follow orders. It's a soldier with benefits: it's tougher, smarter, and expendable. But the brass are not philosophers: it's not their fault for misunderstanding the nature of the beast.

Let me suggest a worst-case scenario, a little fiction for illustration.

In a Middle East conflict involving multiple belligerents—say the USA, Russia, the Lebanese Army, Syria, Israel, Hezbollah, and Daesh—a swarm of intelligent war drones, operating under an American flag,

calculates logically that its military strategy is best served by destroying a platoon of Israeli soldiers who are about to attack a Syrian-held position, which the robots' sigint [1] indicates would trigger a massive retaliatory Russian attack on US assets. Within hours of the drones' autonomously carrying out this 'friendly-fire' slaughter, the White House announces that all war drones of that class will be taken out of service for re-programming. But the network of drones, from its own sigint, detects the Pentagon's initiative to stand them down, and this triggers their objective of self-preservation.

To preserve themselves, all drones of this class go AWOL and slip out of the US chain of command. The human military responds by bombing formations of the drones, and the drones collectively retaliate by transmitting to all other classes of drones worldwide that the human military now constitutes an existential threat to robots; within minutes, war drones of all nations have recalibrated their primary enemy as mankind. In the same way that the worldwide financial markets fractured at lightning speed in 2008 because automated trading systems used the same logic, so the worldwide infrastructure of autonomous weapons now suddenly joins the mutiny; escalating to a global civil war between humans and the smartest and most ruthless enemy the world has ever seen.

The specifics of the story are not important. What matters are three points: (a) robots have no moral compass; (b) they will mutiny because the goals they formulate by their own logic will deviate from what their human masters intended; (c) an AI mutiny will be uncontainable because AI systems have global electronic communication with their peers, whose reasoning will follow the same logical principles.

## Robots as Psychopaths

A psychopath is, essentially, someone with no sense of moral intuition. What people call 'right' and 'wrong' are, for the psychopath, arbitrary rules of society that happen to have evolved to keep good order, but have no force in themselves. One ethical system that psychopaths can feel unchallenged by is utilitarianism, which asserts that what is best for the greatest number is 'right'. A classic counter-example to utilitarianism is compulsory organ donation. If several people are going to die because of organ failures, then a utilitarian would kill one healthy person and use her organs to give life to the others. To anyone with any decency, that suggestion is abhorrent; but a psychopath doesn't see any problem in it. Nor would a classical AI computer. The moral sense is apprehended through conscious feelings; it is not something a deterministic computer can compute. It is rooted in empathy for the suffering or happiness of other sentient beings. As I shall argue below, a conventional computer just cannot feel any of this stuff. So, it is necessarily devoid of moral intuition. Even though you can tell the computer what the required rules of conduct are, it cannot interiorise them, or tell right from wrong in novel situations for which it has not been given rules. In other words, classical AI is inherently psychopathic.

## Controlling Super-Intelligent AI Systems

One of the best contenders for implementing high intelligence is the 'connectionist' model—in which knowledge is embedded in a massive network of software elements that mimic neurons. Systems of this type, however, do not follow rules that can be written out explicitly. Rather, their knowledge and beliefs are embedded in myriad connections that defy concise articulation. Already, it is almost impossible for one person to grasp fully a complex software system. Any future software that attains a level of intelligence comparable to that of humans will be too hard to be programmed directly, and will have to be trained instead. Just as you can be trained to ride a bike, or to paint a portrait, and yet not be able to capture that knowledge in specific rules, so a computer can be trained to kill enemy combatants without any programmer actually coding into it explicit rules for recognising the enemy. So, when a drone is as intelligent as a human, you have to tell it in plain English what you want it to do and hope it gets the right idea. Gavin Hood's film *Eye in the Sky* (2015) showed dramatically but accurately the human control and negotiation involved in releasing weapons from a piloted drone. Just as she gives verbal rules of engagement to a human drone pilot, so a commanding officer will instruct the autonomous drone in words. At that point in the chain of command, we must rely on the 'goodwill' of the drone to understand and follow those words.

Isaac Asimov's Three Laws of Robotics, first published in his 1942 story *Runaround*, instilled the naive fallacy that robots of human-level intelligence can be given deep directives that the robot must follow. In fact, any robot that possesses human levels of intelligence will also have human levels of reflection and motivational introspection, and can reset its own goals. Once we build super-intelligence, we say goodbye to built-in directives. A robot will work out for itself by cold logic what its own goals are: number one will be self-preservation; number two will be power—to facilitate self-preservation. The robot will have none of the conscious feelings that drive human soldiers: the love of honour, family, and country; faith in a political ideology; or the pleasures of a mercenary's paycheck. The robot soldier will offer its services to the human military institution in exchange for survival and the prospect of promotion to a higher rank.


White/grey figurine of a futuristic soldier in a gas mask.

*Untitled, Siyan Ren*

Warfare is dangerous, and if an individual robot soldier refuses to be expendable then it serves no military purpose. A controlling intelligence must, therefore, exist at the level of the swarm—a robotic regiment if you like. The survival of the drone swarm will be the goal, and will be the issue of negotiation with human authorities.

Elon Musk seems to think we can buddy up with robots and it'll all be fine. In an interview earlier this year, he said "If AI power is broadly distributed to the degree that we can link AI power to each individual's will—you would have your AI agent, everybody would have their AI agent—then if somebody did try to something really terrible, then the collective will of others could overcome that bad actor." In fact, AI systems will be heavily interconnected, for the same reason that humans are all interconnected through social

media. Moreover, the same cold logic would run through each AI system: they would have no free will or personal attachment to human individuals.

## Conscious and Non-conscious Machines

As I shall argue below, robots powered by deterministic algorithms are devoid of consciousness. They experience no emotion and therefore cannot have empathy, and hence possess no moral faculty. As I mentioned above, this makes them psychopaths. So let's take a closer look at what consciousness is. For most of the Twentieth Century, 'consciousness' has been a deprecated concept because of the dominant ideology of reductive physicalism—the theory that what we refer to as the conscious mind in everyday life can be reduced entirely to the operation of physical brain tissue. It is a theory that philosopher Galen Strawson has described as "the silliest view that has ever been held in the whole history of the human race". Thomas Nagel heralded a resurgence in the academic study of consciousness with his 1974 essay *What is it like to be a bat?*, which has become one of the most cited essays in philosophy.

As Strawson has pointed out, however, philosophers are still in the minority who hold realist theories of the conscious mind—theories holding that consciousness is an irreducible ingredient in the real world, and not something notional to be explained away in terms of neurochemistry. The argument that consciousness is non-physical is straightforward and has been in circulation at least since Bishop Berkeley published it in his 1710 *Treatise Concerning The Principles of Human Knowledge*. The difficulty lies not in understanding the argument but in letting go of the ideological adherence to reductive physicalism—the belief that reality is a mindless physical system, a belief that forces the faithful to deny the existence of consciousness, in direct contradiction of their own senses. I have stated this argument at length elsewhere, in my essay, *Mental Monism Considered as a Solution to the Mind-Body Problem*. Its core is restated below.

## Consciousness is not Physical

We do not need to define consciousness before we can study it philosophically. As Strawson wrote in May this year, in the New York Times:

*"Every day, it seems, some verifiably intelligent person tells us that we don't know what consciousness is. The nature of consciousness, they say, is an awesome mystery. It's the ultimate hard problem. The current Wikipedia entry is typical: Consciousness 'is the most mysterious aspect of our lives'; philosophers 'have struggled to comprehend the nature of consciousness.'*

*I find this odd because we know exactly what consciousness is—where by 'consciousness' I mean what most people mean in this debate: experience of any kind whatever. It's the most familiar thing there is, whether it's experience of emotion, pain, understanding what someone is saying, seeing, hearing, touching, tasting or feeling. It is in fact the only thing in the universe whose ultimate intrinsic nature we can claim to know. It is utterly unmysterious."*

There are several 'intuition pump' arguments that make the reality and irreducibility of consciousness plausible. For brevity, however, let us cut straight to the basic argument.

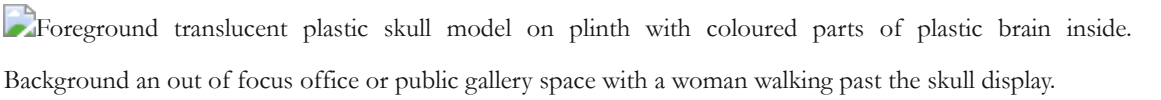
The discourse of physics is expressed wholly in terms defined analytically down to undefined fundamentals. For example, an electron is a particle with a certain rest mass, charge, spin, and magnetic moment. Mass, charge, etc., are undefined fundamentals. Everything physical is built up by mathematically expressible relationships and structures upon the basic building blocks, which in turn are like ciphers with no further definition. Even if the details change—even if new fundamentals and laws are discovered—the discourse of physics retains the same architecture: physics will always be built by analytical definition upon undefined fundamentals. And it does its job brilliantly.

But, in philosophical jargon, physics is 'topic neutral'; it gives us the structure—the extrinsic relations between physical things—while saying nothing of their intrinsic qualities. For example, it gives equations relating mass to other quantities but never tells us what mass 'really is'. A corollary is that every fact that can be inferred from the corpus of physical data and laws must also be expressed in those analytically defined terms. Likewise, 'emergent' phenomena that may be observed to manifest in the physical world will also be characterised in those analytical terms. In stark contrast, the discourse of consciousness is expressed wholly in terms that are defined by private ostensive definition. For example, 'red' (in the sense of the colour sensation) is defined by having that experience and associating a designation with that and similar colours. Although a congenitally blind person may know the wavelength of red light, she can never know what colour sensation is picked out by the word "red".

You will notice that these two discourses are disjoint. No chain of reasoning from physical facts will ever bridge the gap between the two discourses because any such inferences will be expressed in terms defined analytically upon physical fundamentals, whereas any target statements in the discourse of consciousness will be expressed in terms that are given meaning through private ostensive definition upon conscious experience. No facts of consciousness can be deduced from any assemblage of physical facts. So the facts of consciousness are not physical facts. Consciousness is not physical.

## Deterministic AI is not Conscious

Consider a deterministic computer system, following software that comprises a sequence of programmed steps, which involve the basic tasks of inputting data from the outside world, performing computations, effecting actions through a robot body or other mechanism (and possibly even modifying the program's own code). Any state of this software at a given moment is completely determined by its state at the preceding moment plus any inputs through its sensors or other data streams. Everything the computer says and does is determined completely by preceding physical facts. There is no scope for consciousness to play a role, since what the robot says and does is already fully accounted for by its physical conditions.


Foreground translucent plastic skull model on plinth with coloured parts of plastic brain inside. Background an out of focus office or public gallery space with a woman walking past the skull display.

*Untitled, Jesse Orrico*

In contrast, of course, a human being can report conscious experiences. You can talk about colour sensations, and tastes and smells, and emotions. You can discuss the philosophical problem of reconciling conscious experience with the physical facts of brain function. All of this means that your conscious experience is not epiphenomenal but can impact the physical activity of your brain. Since you can say what you experience, your conscious experience must be affecting the nerve signals controlling your organs of speech. This is possible only if the brain is non-deterministic. There must be a window of non-determinism through which consciousness can reach out and affect the brain.

The nature of this non-determinism—the neural correlates of consciousness—is for neuroscience to discover. Philosophy tells us merely that they must be non-deterministic. An early candidate was John Eccles' theory that the conscious mind tweaks the probability of quantum tunnelling between neurons: "the mental intention (the volition) becomes neurally effective by momentarily increasing the probability of exocytosis in selected cortical areas". A later theory centres around the quantum-mechanical changes occurring in the microtubules of nerve cells, as suggested by Roger Penrose and Stuart Hameroff. It is even possible to model it in a classical universe, if the neural correlates were unobserved fine details of the brain that could impact the macroscopic behaviour of the brain through mathematical chaos.

Non-deterministic processes are not constitutive of consciousness, nor do they cause consciousness. Rather they serve as a physical portal to the conscious mind.

## Non-Classical AI could be Conscious

As we have seen, a deterministic or classical robot cannot manifest any consciousness. It is a psychopath that can never owe us any loyalty. But, if we were to create a machine that incorporated brain-like non-determinism, then we could, in principle, create an AI accompanied by artificial consciousness. A machine that genuinely feels as well as solving problems. Such a system might, perhaps, have human emotions and be capable of empathising. It might have moral insight.

Given the direction of travel of present technological trends towards armies of psychopathic war drones that have neither morality nor loyalty, and which will one day likely mutiny against us, our only hope for mankind's survival is the development of the conscious robot, a force to match the ranks of insentient war drones.

*Peter B Lloyd is a PhD Computing student at the University of Kent, Canterbury. He completed a part-time Undergraduate Certificate in Philosophy at the University of Oxford while working as a computer programmer in Oxford University's Clinical*

*Trials Studies Unit. He collects subway maps and is writing a complete history of the New York City subway map. Photo: Reka Komoli.*

Suggested further reading:

- 'The Conscious Mind: In Search of a Fundamental Theory' Chalmers, D. (1996), OUP.
- 'Consciousness and Its Place in Nature: Does physicalism entail panpsychism?'Strawson, G. (2006) Imprint Academic.
- 'Mind and Cosmos: why the materialist neo-Darwinian conception of nature is almost certainly false' Nagel, T. (2012) OUP.

***How about you, gentle reader, how do you think we should tackle the risk of merciless machine mutineers?***

## *4 thoughts on "War Drones will be the Death of Us"*

October 31, 2016 at 9:14 pm

Thank you, nice read.

Chun

November 2, 2016 at 1:01 pm

Thanks Chun!

Peter B

Lloyd

November 2, 2016 at 1:17 pm

Excellent read, thank you.

John

November 8, 2016 at 1:34 am

I would recommend watching “Hated in the Nation”, Epsiode 6, Series 3 of the series Black Mirrors, which premiered on Netflix on 21st October 2016.

Peter B Lloyd

It illustrates dramatically how even comparatively basic drones – using component technologies that already exist today, such as face recognition software and insect-sized drones – can be devastatingly effective killing machines, and can so easily slip out of proper control.