Blog of the APA

A Neurophilosophy of Autonomous Weapons and Warfare

By Nayef Al-Rodhan

August 10, 2020

This is post nine in a short-term series by Prof. Nayef Al-Rodhan titled "Neurophilosophy of Governance, Power and Transformative Innovations." This series provides neurophilosophical perspectives and multi-disciplinary analyses on topics related to power and political institutions, as well as on a series of contemporary transformative technologies and their disruptive nature. The goal is to inspire innovative intellectual reflections and to advance novel policy considerations.

A starting point in discussions of autonomy in weapons systems is the comparison between the consequences of errors in robotic or semi-automatic systems and similar wrongdoings in a human soldier. Tragic human errors often occur in warfare but such incidents can usually be traced back to one or more individuals who are subsequently held responsible. 'Responsibility' for "killer robots" is more complicated. The issue of accountability is key in this debate, but in addition to that, there is also the highly disputable issue of reliability and deciding at what point an autonomous weapon system could be deemed fully able to cope with real-life situations. Human soldiers *cannot* realistically be trained to respond to all possible scenarios either but military training endows them with unique experience and agility to guide their decisions even in unpredictable situations.

A neurophilosphical perspective on autonomous weapons and their use in warfare goes even further in explaining why the acceptance of autonomous weapons is so profoundly antagonistic to human nature.

Autonomous weapons – contours of the debate

The US Department of Defence Directive 3000.09 covering "Autonomy in weapons systems" from 2012 (updated in 2017) defines autonomous weapons systems as "a weapon system that, once activated, can select and engage targets without further intervention by a human operator". The same document states that as a matter of policy, "systems will go through rigorous hardware and software verification and validation and realistic system developmental and operational test and evaluation". Such points have been anything but reassuring to those against autonomous weapons.

Fully autonomous weapons are not yet fully deployed in warfare but robots with extensive autonomous capabilities have assisted military operations for years, decades even, particularly in roles of surveillance, reconnaissance, for detonating IEDs, or surveying damage from biochemical weapons. Most versions of these weapons systems, which include a wide range of capabilities, from cruise missiles, torpedoes, unmanned underwater vehicles, aerial robots, to space robots, still require a significant degree of human supervision. However, as sensor technology and AI continue to improve, the need for supervision would gradually decrease. This was one of the underlying goals of Pentagon's *Project Maven*, which employed Google's AI software, signaling a clear interest to use AI in future weaponry, in particular by using AI to identify objects in drone footage. Google withdrew from the Project after thousands of employs signed a petition citing ethical concerns, and after several engineers refused to build one of the security tools for the Project. Moreover, the dissenting employees asked the company to swear off any further involvement in military work, which prompted a commitment from the company's top leadership to develop responsible AI, without refusing all defense work altogether.

Such calls echo the messages of numerous campaigns from civil society and scientists to ban killer robots on the grounds that they can make fatal errors while evading a clear chain of responsibility and accountability.

A core conviction of those in favor of a complete ban on autonomous weapons is that machines, no matter how advanced, can never comply with the legal and moral requirements of the laws of war. The mere fact of 'autonomy' in a machine is hardly comparable to autonomy in a human being, which is a basis for them to act as free moral agents. The concept of autonomy has been long debated in philosophy and it was central to Kant's idea of morality because it recognized our capacity to exercise free will and that the 'authority' binding us to our principles was not separate from our will. In machines, that autonomy means weapons are expected to adapt to unforeseen circumstances in their environment. The fact that they will behave unpredictably does not mean 'randomness', as Robert Sparrow notes, because randomness in itself is not a ground for a claim to autonomy. The robot's decisions will follow reasons "responsive to the system itself" but it remains unclear at this point how developed that autonomy could ever be.

In any case, as Sparrow notes, if a weapon is to be called 'autonomous' and therefore it acts as an autonomous agent, then it cannot hold anyone else responsible for its actions – this raises a big dilemma because, as outlined in the beginning, no machine can be legally held accountable for its actions, which makes any hope of prosecution moot. Or, having weapons that can make errors frequently and do so in a way that it makes it impossible to identify responsibility is contrary to the rules of *jus in bello*.

Fundamental principles of the laws of armed conflict include the principle of *distinction*, which requires combatants to distinguish between combatants and civilians and between military and civilian objects, and the principle of *proportionality* in the use of force – which precludes exceeding a threshold of damage beyond what is needed for military advantage. There are other, less 'codified' considerations for banning autonomous weapons in warfare and this include, as a matter of principle, the fact that it is morally wrong to allow a machine to be in charge of making decisions of life and death of another human being (a position upheld, among others by the Holy See, as expressed at a meeting on lethal autonomous weapons at the UN in 2014).

Decision-making and the responsibility gap

A sophisticated autonomous weapon's judgments of distinction and proportionality would be probabilistic: this means that programmers would need to attach value to certain targets, humans and objects and then make probabilistic assessments accounting for contextual factors.. Humans too make probabilistic decisions and by weighing in contextual factors; however, (and as the argument against autonomous weapons goes), there are irreplaceable human elements in warfare such as compassion and empathy for other human beings, which can change the course of certain actions. For a machine to replace humans in warfare, the former should be able to uphold the fundamental requirements of the laws of war, as a basic requirement. Even then, many argue, these weapons remain *mala in se* simply because they take humans out of the loop and therefore remove the only 'true moral agent' endowed with a conscience from crucial decisions – as flawed as those human decisions may be. Humans, unlike robots, can think about consequences beyond strictly deontological calculations.

This feeds into the aforementioned argument against autonomous weapons, which is the complex issue of accountability: should a machine make errors (and especially if this could go as far as committing war crimes), who is to be held responsible: the programmer, the manufacturer, the commander that employed the robot on the battlefield? The 'responsibility gap' problem will make it close to impossible to decide who to hold responsible in cases of fatal accidents. The calls for 'meaningful human control' over future weapon systems, issued by states and multilateral organizations alike, has only exposed, however, the theoretical and practical complexity of the matter. Defining what "meaningful" means is not clear-cut, and neither is the meaning of keeping humans "in the loop", which does not seem be a satisfactory condition in itself as simply being in the loop does not mean being in control of a military decision. Indeed, the problem of responsibility is critical and to date, yet to be solved.

We might ask at this point why the idea of developing autonomous weapons systems has endured for so long, and continues to remain a powerful one. The reasons are pragmatic and generally centered on the robots' military utility, which would replace humans in "dull, dirty, and dangerous jobs" (the so-called 3Ds of robotization). Robots could, among others, replace the human soldiers in highly dangerous missions, such as tunneling through dark caves in search of enemies, clearing roads and waters from explosive devices, patrolling urban streets in rife with sniper fire, or surveying damage from biochemical weapons. As the argument goes, robots would be unaffected by fatigue, and the attendant effects on cognitive abilities, or by emotions, which have long swayed soldiers in warfare, or from the debilitating stress that has often made soldiers overstep the rules of armed conflict.

These ideas may be very appealing, yet that does not take away the serious ethical and legal problems highlighted above. While increased robotization for certain tasks and functions in war can be welcome, any autonomous system equipped with lethal capabilities will meet with serious and persistent concerns from ethicists and lawyers.

A neurophilosphical perspective on autonomous weapons highlights, further, why their acceptability (in their current form, at least) is antagonistic to defining elements of our nature.

Neurophilosophical perspectives on autonomous weapons

1. Human nature and accountability

Previous sections emphasized the critical aspects of the 'responsibility gap' and accountability. As Sparrow underlined, the principle that 'we must be able to identify those responsible for deaths in war (…) is a necessary condition of the respect for persons that is at the heart of Kantian, and other deontological, ethics.' Evidence from neuroscience allows us to discuss these considerations in the context of human nature.

With insights from neuroscience, and as elaborated in further detail in previous posts, I described human nature as *emotional, amoral* and *egoistic*. Humans are far more *emotional* than rational, a conclusion demonstrated by neuroscientific research in recent decades – and an uncomfortable reality if we consider just how much 'emotions' have been considered a hindrance to rational decisions in the history of philosophy. In fact, emotional processing in the brain is deeply connected with a host of other cognitive processes, learning, and decision-making. *Amorality* is another defining elements of human nature. Contrary to theories that consider humans to have been born either good or bad, this is hardly supported by neuroscientific evidence. We are neither innately moral, nor immoral, but rather amoral – meaning that our moral compass is largely "written upon" in the course of our existence and depending on circumstances in our environment. We do possess, however, a set of inborn predilections, the most powerful of which is the predisposition for survival that will guide us to always choose those actions that maximize our chances of survival. We are in this sense a *predisposed tabula rasa* – an amendment to John Locke's theory of tabula rasa. Our *egoism* largely derives from this basic predisposition for survival, which is a basic form of egoism (of the self, or of our kin). Against the backdrop of these defining elements in our nature (emotionality, amorality and egoism) governance models must include considerations of human dignity.

Dignity has been absent from many indexes of good governance but it is fundamental (even more so than 'freedom') to ensuring that the best in our nature is allowed to thrive. I define dignity not in a reductionist sense as the absence of humiliation but as a more holistic set of nine dignity needs: *reason, security, human rights, accountability, transparency, justice, opportunity, innovation and inclusiveness*. In a previous post, I discussed how each of the three defining traits of human nature can be paired with three dignity needs.

I want to refer here in particular, and germane to this discussion of autonomous weapons, to **accountability** and **justice**, dignity needs that are critical in the context of man's *amoral* character. Accountability is paramount to promote pro-social behavior and to strengthen trust in any judicial system. Our preoccupation with accountability is deeply connected to our interests in **justice** and fairness, which are strongly represented in the brain, across several brain regions that are concerned with social decision-making.

Considerations related to righteousness and justice permeate interindividual behavior. Several brain regions are activated when we make moral choices or when considering moral dilemmas, and moral sensitivity is deeply implicated in other cognitive and emotional processes. Furthermore, one region known to be engaged in moral judgment, the ventromedial prefrontal cortex plays an important role in mediating emotions during moral processing, as well as in adherence to social norms, among others; lesions in this brain region have been associated with higher likelihood of utilitarian responses to hard moral dilemmas. Another region, the orbitofrontal cortex has been associated with other 'facets' of morality, such as representations of reward and punishment. The anterior cingulate cortex has been shown to be involved in error detection, and the medial frontal gyrus is involved in some social functions related to moral judgment.

The accountability issues that beset the use of lethal autonomous weapon systems must also be understood in the context of these defining traits of human nature. The deployment of autonomous weapons, no matter how well justified and utilitarian, cannot be reconciled with our nature as long as the risks of lack of accountability and injustice remains ever-present.

2. Military ethics and the human soldier

There are, additionally, elements pertaining to the domain of military ethics in particular that must be accounted for as well. The use of autonomous weapons systems, especially as human soldiers also remain critical actors in warfare bring about further ethical problems – some of which are also connected to the concerns about justice and accountability described above.

The notion of honor and courage within the warrior ethos distinguishes the military professionals from civilians, and is a reality carved within the military community praised by Clausewitz as he discussed this profession as a pursuit different from others that 'occupy the life of man'. Western militaries employ *hybrid ethical codes* that draw inspiration from Aristotelian virtue ethics, and deontological ethics. Virtue ethics stresses a. the contextual relativity of virtues; b. actor relativity; c. the emphasis on character formation. In the unpredictability of war, virtue ethics is meant to guide the soldier through morally challenging and unexpected situations. Aristotle had called this "phronesis": the ability to apply the appropriate virtue at the right time. Therefore, in addition to deontological ethics, which dictates abiding by absolute rules (e.g. laws of armed conflict), military training has a role in carving the soldier's character by developing his/her courage and ability to apply virtue-guided judgment.

It is hard to imagine that the execution of highly complex military tasks, particularly in an urban warfare setting, can be delegated to robots. Even if we were to leave aside the premise that humans are uniquely able to make critical moral judgments, the consideration of fairness may play out here too. Should soldiers fight alongside autonomous weapons, their sense of belonging to the military as a community will be severely damaged as they will share their duties with robots that never underwent the training or sacrifices they were exposed to. They may feel outranked in their status and this will have profound consequences for the social and moral fabric of the military community. Even if an all-robotic scenario is plausible for certain operations, war is always going to remain a human affair, where human soldiers will remain, in varying degrees and roles, indispensable. In that case, while perhaps grateful to be able to avoid some life-threatening situations, soldiers' missions and sense of worth in the community will be strongly challenged if their roles are rendered less meaningful.

3. 'Moral robots' – whose morality will they inherit?

A final, and more futuristic, point about autonomous weapons systems regards a scenario in which robots too will develop moral competences, be able to distinguish combatants from non-combatants with the same accuracy that a human soldier can and make fully independent moral judgments. We are now very far from that moment, yet developments such as in the field of neuromorphic chip technology, which work to create chips that emulate the human neuronal networks, make the prospect of autonomous robots more real. This would effectively mark the transition from top-down to bottom-up acquisition of morality, therefore from a set of moral rules programmed into the system by engineer, ethicist or software engineer, to a type of morality acquired by robots through a social process of learning and engagement with the world. Yet, this does not necessarily mean these robots will be moral by default. A fundamental question in this scenario is whose morality would robots inherit? Would they mirror the traits of human nature, i.e. *emotionality, amorality* and *egoism*? This also imply they would develop a strong interest in their own self-preservation and thus be less inclined to sacrifice their existence in war. The absence of real autonomy, which also means a capacity for moral judgment and accountability, is something that is currently held against autonomous weapons.

But would the development of bottom-up moral competence solve the ethical dilemmas surrounding autonomous weapons? In other words, is the critique against lethal autonomous weapons simply about the lack of a sufficient level of autonomy, which would imply that the existence of full autonomy would solve the problem?

A short answer is no. As Christian Enemark noted in his work on armed drones, "ethics is […] constitutive of the practice of war as a form of violence that is (…) morally distinguishable from other forms". However, a future of fully autonomous weapons in warfare will be defined by immense levels of asymmetry – it is indeed highly improbable that all countries will possess such advanced systems – and this will shatter current understandings of *jus ad bellum* and *jus in bello*.

Conclusions

Some of the malfunctions in autonomous systems that came to public light in recent years are an indication why the moral and legal acceptance of autonomous weapons could not happen for a very long time. For example, in 2016, the sensors of a self-driving car failed to detect a big white truck against a bright spring sky, mistakenly classifying the truck for a portion of the sky and killing the passenger of the car. Such technological mishaps – it would appear – is what still prevents us from sending killer robots off to war.

By that logic, achieving full autonomy and advanced robot capabilities to conscientiously apply the laws of armed conflict, is what still separates us from the moment of full deployment of autonomous weapons systems. However, as outlined in this piece, the range of moral arguments against autonomous weapons is far more complex. A neurophilosophical account further clarifies why their acceptability is so difficult and deeply rooted, among others, in considerations of justice and fairness – both of which are deeply embedded in human neuroanatomy and neurochemistry.

Reassurances and 'ethical codes' from industry and policy makers remain, in lights of these considerations, insufficient for the full acceptability of autonomous weapons systems. A very recent example is from February 2020, when the US DoD adopted a set of ethical principles for artificial intelligence. Though not explicitly referring to fully autonomous weapons, this is a welcome step to commit to develop AI that is 1."responsible", 2."equitable", 3. "traceable", 4. "reliable" and 5. "governable". Applying these principles in practice will depend both on institutional commitment and technological advancements (indeed, committing to having 'traceable' systems is meaningless if the technology is not developed enough to allow it). Crucially, however, a neurophilosophical account of human nature must be part of the debate. The acceptability of autonomous weapons can be achieved only when (and if) the tensions between our *emotional, amoral, egoism,* and *human dignity* needs will have been fully addressed.

Nayef Al-Rodhan

Prof. Nayef Al-Rodhan (@SustainHistory) is a Neuroscientist, Philosopher and Geostrategist. He is an Honorary Fellow at St Antony's College, University of Oxford, and Senior Fellow and Head of the Geopolitics and Global Futures Programme at the Geneva Centre for Security Policy, Geneva, Switzerland. Through many innovative books and articles, he has made significant conceptual contributions to the application of the field of neurophilosophy to human nature, history, contemporary geopolitics, international relations, cultural studies, future studies, and war and peace.