On the Autonomy and Threat of "Killer Robots"

Jean-Gabriel Ganascia, Catherine Tessier, Thomas M Powers

On the autonomy and threat of “killer robots”

Jean-Gabriel Ganascia[1], *Catherine Tessier*[2], *Thomas M. Powers*[3]

1. Sorbonne University; Member of the *Institut Universitaire de France*, Chairman of the CNRS Ethical Committee

2. ONERA Aerospace Lab, France; Information Processing and Systems Department

3. University of Delaware; Department of Philosophy and Center for Science, Ethics & Public Policy

Introduction

In the past, renowned scientists such as Albert Einstein and Bertrand Russell publicly engaged, with courage and determination, the existential threat of nuclear weapons. In more recent times, scientists, industrialists, and business leaders have called on states to institute a ban on what are — in the popular imagination — "killer robots." In technical terms, they are objecting to LAWS (Lethal Autonomous Weapons Systems), and their posture seems similar to their earlier, courageous counterparts. During the 2015 International Joint Conference on Artificial Intelligence (IJCAI) — which is the premier international conference of artificial intelligence — some researchers in the field of AI announced an open letter warning of a new AI arms race and proposing a ban on offensive lethal autonomous systems. To date, this letter has been signed by more than 3,700 researchers and by more than 20,000 others, including (of note) Elon Musk, Noam Chomsky, Steve Wozniak, and Stephen Hawking.

In the summer of 2017, at the most recent IJCAI held in Melbourne, Australia, another open letter was presented, signed by the heads of many companies in the fields of robotics and information technologies, among whom Elon Musk was very active. This second letter urged the United Nations to resume its work towards a ban on autonomous weapons, which had been suspended for budgetary reasons.

It is no doubt incumbent on every enlightened person, and in particular on every scientist, to do everything possible to ensure that the industrialized states give up the idea of embarking on yet another mad arms race, the outcome of which might escape human control. This seems obvious, especially since, according to the authors of these two open letters, we would be at the dawn of a third revolution in the art of war, after gunpowder and the atomic bomb.

If these positions appear praiseworthy at first, should we not also wonder about the actual threats of these lethal autonomous weapon systems? To remain generous and sensitive to great humanitarian causes, should we not also remain rational and maintain our critical sensibilities? Indeed, even though considerable ethical problems arise in the evolution of armaments — from landmines to drones, and recently to the massive exploitation of digitized information and electronic warfare — it appears on reflection that this third revolution in the art of war is very obscure. Where the first two

revolutions delivered considerable increases in firepower, we find here an evolution of a very different order.

Moreover, the so-called “killer robots” that have been the targets of three years of numerous press articles, open letters and debates, seem to be condemned by sensational and anxiety-laced arguments, mostly to the exclusion of scientific and technical ones. The term “killer robot” suggests a robot that would be driven by the *intention* of killing and would even be *conscious* of that intention, which at this stage in the science does not make sense to attribute to a machine — even one that has been designed for destroying, neutralizing or killing. For instance, one does not speak of a “killer missile,” when it happens that a missile kills someone. "Killer robot" is a term that is deployed for rhetorical effect, that works to hinder ethical discussion, and that aims at manipulating the general public. Do the conclusions of these arguments also hold against "killing robots"? Is there an unavoidable technological path from designing "killing robots" to deploying "killer robots"?

To get a better understanding of these questions, we aim here to put forward a detailed analysis of the 2015 open letter, which was one of the first public manifestations of the desire to ban LAWS. Our reservations concerning the declarations that this letter contains should help to open the scientific and philosophical debates on the controversial issues that lie at the heart of the matter.

The Argument for a Ban

The 2015 open letter was revealed to journalists and, by extension, to a broad audience during the prestigious IJCAI in Buenos-Aires, Argentina. In its first sentence, the letter warned that "[a]utonomous weapons select and engage targets without human intervention," and concluded after four short paragraphs by calling for a ban on offensive forms of such weapons. This public announcement had been preceded by an invitation for signatories within the AI scientific community and beyond, including a wider community of researchers, technologists, and business leaders. Many of the most prominent AI and robotics researchers signed it, and outside the AI community many prominent people brought their support to this text. Initially, the renown and humanitarian spirit of the co-signers may have inclined many people to subscribe to their cause. Indeed, the possibility of autonomous weapons that select their targets and engage lethal actions without human intervention appears really terrifying.

However, after a careful reading of the first open letter, and in consideration of the subsequent public statements on the same topics, e.g., the IJCAI 2017 (second) open letter, and video[1] that circulated widely on the web towards the end of 2017, we think a closer analysis of the deployed arguments clearly shows that the letter raises many more questions than it solves. Despite the fame and the scientific renown of the signatories, many statements in the letter seem to be questionable from a scientific point of view. In addition, the text encompasses declarations that are highly disputable and that will certainly be belied, very soon, by upcoming technological developments. These are the reasons why, as scientists and experts in the field, it seems incumbent upon us to scrutinize the claims that these public announcements contain and to re-open the debate. We are not disparaging the humanitarian aims of the authors of the letter; we

[1] https://www.youtube.com/watch?v=9CO6M2HsoIA&feature=youtu.be

do however want to look more closely at the science and the ethics of this issue. Even though we share the same feeling of unease that has likely motivated the authors and the signatories of these open letters, we want to bring into focus where, we believe, the scientific case is lacking for the normative conclusion they draw.

For ease of reference, the content of the 2015 Open Letter has been appended to this article, with numbered lines added to facilitate comparison between our text and theirs.

The first paragraph (l. 10-17) describes recent advances in artificial intelligence that will usher in a new generation of weapons that qualify as autonomous because they *"select and engage targets without human intervention."* These weapons will possibly be deployed *"within years, not decades"* and will constitute *"the third revolution in warfare, after gunpowder and nuclear arms."* The next paragraph (l. 18-33) explains why a military artificial intelligence arms race would not be beneficial for humanity. The two main arguments are, first, that *"if any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable"* and second, as a consequence, "*autonomous weapons will become the Kalashnikovs of tomorrow,"* i.e. they will become ubiquitous because they will be cheap to produce and distribution will flow easily from states to non-state actors. In addition, this paragraph warns that autonomous weapons are *"ideal"* for dirty wars, i.e. *"assassinations, destabilizing nations, subduing populations and selectively killing a particular ethnic group."* The third paragraph (l. 34-40) draws a parallel between autonomous weapons and biological or chemical weapons, the development of which most scientists have rightly shunned. AI researchers, it is implied, would *"tarnish their field"* by developing AI weapons. Finally, the last paragraph (l. 41-44) summarizes the content of the letter and then calls for a ban on offensive autonomous weapons.

Our perplexity comes from these four aspects of the general argument, as developed in the letter:

1. The notion of "autonomous weapon" that motivates the letter is obscure; its novelty and what distinguishes it from AI weapons in general are sources of confusion. At least this much is certain: not all AI weapons are autonomous, according to the definition given by the authors (selecting and engaging targets without human intervention.) Contrary to what is claimed, the technical feasibility of autonomous weapons deployment in the near future is far from obvious.
2. Despite the dramatic illustrations given in the letter, and repeated in the video to which we referred above, the specific noxiousness of autonomous weapons that makes them *"ideal"* for dirty military actions and that differentiates them from current weapons is not obvious from a technical point of view.
3. The analogy between the current attitude of AI scientists faced with the development of autonomous weapons and the past attitude of scientists faced with the development of chemical and biological weapons is far from clear. Besides, the parallel between the supposed outbreak of autonomous weapons in contemporary military theaters and the advent of gunpowder or nuclear bombs in warfare is highly debatable.
4. Lastly, the ban on offensive autonomous weapons is not new and is already being discussed by military leaders themselves, which makes this declaration somewhat irrelevant.

The remainder of this article is dedicated to a deeper analysis of the four points above.

Autonomous Weapons

What exactly is the notion of “autonomous weapon” to which the letter refers? Autonomy is the capability for a machine to function independently of another agent (human, other machine) exhibiting non-trivial behaviors in complex, dynamic, unpredictable environments [1]. The autonomy of a weapon system would involve sensors to assist in automated decisions and machine actions that are calculated without human intervention. Understood in this way, autonomous weapons have already existed for some time, as exemplified by a laser-guided missile that “hangs” a target.

The current drones that are operated and controlled manually at more than 3000 km from their objectives use such autonomous missiles. If this were the meaning of “autonomous weapons” in this letter, the notion would correspond only to a continuous progression in military techniques. In other words, this would just be an augmentation in the distance between the “soldier” (or, more precisely the operator) and its target. In this respect, among a bow-and-arrow, a musket, a gun, a canon, a bomber and a drone, there is just a difference in the order of magnitude of the arms’ ranges. However, the text of the open letter does not say this, but rather claims that (l. 10) *[a]utonomous weapons select and engage targets without human intervention.* The question then is not about the range of action but about the “logical” nature of the weapon: until now, and for centuries, a human soldier aimed at the target before firing, while in the future, with autonomous weapons, the target will be abstractly specified in advance. In other words, the mode of designating the target changes. While up to now, the objective, i.e., the target, was primarily an index on which the human aimed, in the near future it will just become an abstract symbol designated by a predefined rule. Since no human is involved in triggering the lethal action, this evolution of warfare seems terrifying, which would justify the concerns of the open letter.

Let us note that the concept of "autonomy" is problematic, firstly because various stakeholders (among them, scientists) give the term multiple meanings [2] [3]. An "autonomous weapon" can thus designate a machine that reacts automatically to certain predefined signals, that optimizes its trajectory to neutralize a target for which it has automatically recognized a predefined signature, or that automatically searches for a predefined target in a given area. Rather than speaking of "autonomous weapons," it seems more relevant to study which functions are or could be automated, which is to say, delegated to computer programs. Further, we should want to understand the limitations of this delegation, in the context of a sharing of authority (or control) with a human operator, which sharing may vary during the mission.

Guidance and navigation functions have been automated for a long time (e.g. automatic piloting) and have not raised significant questions. These are non-critical operational functions. But automatic identification and targeting are more sensitive functions. Existing weapons have target recognition capabilities based on predefined models (or signatures): the recognition software matches the signals received by the sensors (radar signals, images, etc.) with its signature database. This recognition generally concerns large objects that are "easy" to recognize (radars, airbases, tanks, missile batteries). But the software is unable to assess the situation around these objects – for example, the presence of civilians. Targeting is carried out under human supervision, before and/or during the course of the mission.

Ineluctability

The authors seem to suggest that this evolution is ineluctable because, if specification of abstract criteria and construction of the implementing technology is cheaper and faster than recruiting and training soldiers, and assuming that modern armies have the financial and technical wherewithal to make these weapons, then autonomous weapons will eventually predominate. This complicated point deserves some more in-depth analysis, since the definition of the *criteria* to which the open letter refers appears sometimes very problematic, despite the progress of AI and machine learning techniques. Many problems remain to be solved. For instance, how will the technology differentiate enemies from friends in asymmetric wars, where the soldiers don't wear uniforms? More generally, when humans are not able, on the basis of a given set of information, to discriminate cases that meet criteria from cases that don't, how will machines do better? If humans cannot discern, from photos, which are the child soldiers and which are children playing war, it is illusory to hope to build a machine that automatically learns these criteria, on the basis of the same set of information. Will algorithms be able to recognize a particular individual from their facial features, a foe from their military uniform, a person carrying a gun, a member of a particular group, a citizen of a particular country whose passport will be read from a remote device? It will be impossible to build a training set.

In recognition of these remaining problems, it seems that the supposed ineluctability of the evolution that would spring from the AI state of the art is debatable, and certainly not "*feasible within a few years*" as the letter claims. It would have been more helpful had the authors of the letter elaborated on what precisely will be feasible in the near future, especially as far as automated situation assessment is concerned. The assertion that full-blown autonomous weapons are right around the corner would then have been placed in context.

On the Formal Specifications of Autonomy

Current discussions and controversies focus on the fact that an autonomous weapon would have the ability to recognize complex targets in situations and environments that are themselves complex, and would be able to engage (better than can humans) such targets on the basis of this recognition. Such capabilities would suppose the weapon system:

- to have a formal (i.e. mathematical) description of the possible states of the environment, of the elements of interest in this environment and of the actions to be performed, even though there is no "standard situation" or environment
- to recognize a given state or a given element of interest from sensor data
- to assess whether the actions that are computed respect the principles of humanity (avoid unnecessary harms), discrimination (distinguish military objectives from populations and civilian goods), and proportionality (adequacy between the means implemented and the intended effect) of the International Humanitarian Law (IHL).

Issues of a philosophical and technical nature are related to the ability of the system to automatically "understand" a situation, and in particular to automatically "understand" the intentions of potential targets. Today, weapon system actions are undertaken with human supervision, following a process of assessment of the situation, which seems difficult to formulate mathematically. Indeed, the very notion of agency, when humans

and non-human systems act in concert, is quite complicated and also fraught with legal peril.

Beyond the philosophical and technical aspects, another issue is whether it is ethically acceptable that the decision to kill a human being, who is identified as a target by a machine, can be delegated to this machine. More specifically, with respect to the algorithms of the machine, one must wonder how and by whom the characterization, model and identification of the objects of interest would be set, as well as the selection of *some* pieces of information (to the exclusion of some others) to compute the decision. Moreover, one must wonder who would specify these algorithms and how it would be proven that they comply with international conventions and rules of engagement. And as we indicated above, the accountability issue is central: who should be prosecuted in case of violation of conventions, or misuse? It is our contention that these difficult formal issues will delay (perhaps indefinitely) the advent of the sort of autonomous weapons that the authors so fear.

Finally, it is worth noting that the definition of autonomous weapons (*Autonomous weapons select and engage targets without human intervention* (l. 10)) comes from the 2012 U.S. Department of Defense Directive Number 3000.09 (November 21, 2012. Subject: Autonomy in Weapon Systems). Nevertheless, the authors of the letter have truncated it. As a matter of fact, the complete definition given by the DoD directive is the following: *Autonomous weapon system: a weapon system that, once activated, can select and engage targets without further intervention by a human operator. This includes human-supervised autonomous weapon systems that are designed to allow human operators to override operation of the weapon system, but can select and engage targets without further human input after activation.*

From the DoD directive one learns in particular that (3) "*Autonomous weapon systems may be used to apply non-lethal, non-kinetic force, such as some forms of electronic attack, against materiel targets*" in accordance with DoD Directive 3000.3. Therefore, we should bear in mind that a weapon (in general) should be distinguished from a lethal weapon. Indeed, a weapon system is not necessarily a system that includes lethal devices.

Hence, the proffered, alarming example of what autonomous weapons technology could bring — "*armed quadcopters that can search for and eliminate people meeting certain pre-defined criteria*" (l. 11-12) — seems more fitting for the tabloid press. For this example to be taken seriously, some of those targeting criteria should be made explicit, and current and future technology should be examined as to whether a machine would be able to assign instances to criteria, with no uncertainty, or with less uncertainty than a human assessment. For example, the criterion "target is moving" — for which no AI or autonomy is required — is very different from the criterion "target looks like this sketch and attempts to hide."

Harmfulness

The second paragraph (l. 18-33) is mainly focused on the condemnation of automated weapons.

The Ethics of Robot Soldiers

From the beginning, this paragraph seems intended to measure the costs and benefits of autonomous weapons, but it proceeds too quickly by dismissing debates about the

possible augmentation or diminution of casualties with AI-based weapons. While the arguments for augmentation rely upon the possible multiplication of armed conflicts, the arguments for diminution seem to be based on the position of the roboticist Ronald Arkin [4]. According to Arkin, robot soldiers would be more ethical than human soldiers, because autonomous machines would be able to keep their "blood cold" in any circumstance and to obey the laws of the conduct of a just war. Note that this argument is suspect, because the relevant part of just war laws — the conditions for just conduct or *jus in bellum* — are based on two further principles. As we indicated above, the principle of *discrimination*, according to which soldiers have to be distinguished from civilians, and the principle of *proportionality*, which limits a response to be proportional to the attack, are both crucial to building an ethical robot soldier. Neither discrimination nor proportionality can be easily formalized, so it is unclear how robot soldiers could obey the laws of just war. The problem is that, as mentioned in the previous section, there is no obvious way to extract concrete objective criteria from these two abstract concepts. However, interestingly, the open letter never mentions this formal problem, even though it could help to reinforce its position against autonomous weapons.

Ideal Weapons for Dirty Tasks

The main argument concerning the harmfulness of autonomous weapons is that they *"are ideal for tasks such as assassinations, destabilizing nations, subduing populations and selectively killing a particular ethnic group."* The different harms belonging to this catalogue appear to be highly heterogeneous. What is common to these different goals? Further, the adjective *"ideal"* is particularly obscure. Does it mean that these weapons are perfectly appropriate for the achievement of those dirty tasks? If that is the case, it would have helped to give more details and to show how autonomous weapons would facilitate the work of assailants. Such an elaboration would have been important because, at first glance, there is no evidence that autonomous weapons will be more precise than classical weapons (e.g. drones) for assassination or selective killing of a particular ethnic group. Indeed, it is difficult to imagine how autonomous machines could select, more efficiently than other weapons, the individuals that are to be killed, or discern expeditiously members of human groups, depending on their race, origin or religion. Finally, the underlying premise of the "harmfulness" argument is worth questioning, for it is not clear that those conducting "dirty wars" care much about precision or selectivity. Indeed, this "not caring" may be a central trait of the "dirtiness" of such aggression.

Necessary Distinctions

Underlying the discussion of these loosely-related "dirty" tasks and a possible arms race, there is a confusion between three putative properties of autonomous weapons that, taken one by one, are worth discussing: firepower, precision and diffusion. Despite the reference to gunpowder and nuclear weapons (l. 16-17, 24, 40), there is no direct relation between autonomy of arms and their firepower. Further, it is not any more certain that autonomous weapons would reach their targets more precisely than classical weapons. The series of "drone papers"[2] shows how difficult it is to systematize

[2] A series of papers published by an online publication ("The Intercept") details the drone assassination program of U.S. forces in Afghanistan, Yemen and Somalia. https://theintercept.com/drone-papers/

human targets selection and to automatically gather exact information on individuals by screening big data. Lastly, the argument about the diffusion of autonomous weapons is in contradiction with the supposed specific role of major military powers in autonomous weapon development. More precisely, the problem appears when we consider the following claims:

1. *If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable* (l. 21-23), (which we consider to be probable)
2. *autonomous weapons will become the Kalashnikovs of tomorrow* (l. 24), (which is also possible)

However, even if claims 1 and 2 above are plausible separately, they seem jointly implausible. (By comparison, the development of nuclear weapons *did* start an arms race, but also kept nuclear armaments out of the hands of all but the "nuclear club" of nations) There may even be an antinomy between 1 and 2, because if *only* major military powers would be able to promote scientific programs to develop autonomous weapons, then it is likely that these scientific programs would be too costly to develop for industries, without rich state support, or for poor countries or non-state actors, which means that these arms couldn't so quickly become sufficiently cheap that they would spread throughout all humankind. Some weapons might be more easily replicated, once information technologies have been developed, and military powers could act as pioneers in that respect. However, nowadays, it appears that military industries are not guiding technical development in information technologies, as was the case in the 20th century (at least until the end of the seventies), but that more often the opposite is the case: information technology industries (and dual-purpose technologies) are ahead of the military technologies. Undoubtedly, information technology industries would become prominent in developing autonomous weapons technologies if there were a mass market for autonomous weapons, as the authors of this open letter assume. Lastly, if these technologies were potentially so cheap that they could be spread widely, there would be a strong incentive for the major military powers to keep "a step ahead" to ensure the security of their respective populations.

The paragraph ends with a rather strange sentence (l. 32-33): "*There are many ways in which AI can make battlefields safer for humans, especially civilians, without creating new tools for killing people*." This suggests that AI would benefit defense whereas autonomous weapons would not. Nevertheless, what has been argued previously against autonomous weapons can fit all other AI applications in defense in the same way. Moreover, and to add to the confusion in this claim, the terms *autonomous weapon* (l. 10, 15, 18, 24, 29, 43), *AI weapon* (l. 22, 35) and *AI arms* (l. 21, 31, 42) seem for the authors to be interchangeable or synonymous phrases. Yet equipping a weapon, whether lethal or not, with some AI (e.g. a path planning function) does not necessarily make it autonomous and conversely some forms of autonomy (e.g. an autopilot) may hinge on automation without involving any AI.

Analogies with other Weapons

A third central claim in the general argument concerns military analogies with other weapons: nuclear weapons on the one hand and biological and chemical weapons on the other. All of these parallels are troublesome.

Third Revolution in Warfare

It is announced (l. 15-17) that the development of autonomous weapons would correspond to a third revolution in warfare, after gunpowder and nuclear weapons. Later, the analogy with nuclear weapons is repeated twice (l. 24 and l. 40), in order either to draw connections or to underline differences. Based on our observations above, it does not seem that autonomous weapons will lead to an augmentation in firepower, but instead, to an increase in the distance between the soldier and his/her target. If there is something innovative in autonomous weaponry, it is in *range* rather than *power*. Therefore, it would have been better to compare autonomous weapons with the bow-and-arrow, the musket, or the bomber drone, instead of with weapons for which incidence range is totally heterogeneous.

Parallel with chemical and biological weapons

The third paragraph draws a parallel between autonomous weapons and weapons that have been considered morally repugnant, such as the chemical and biological weapons that scientists don't develop anymore, because they "have no interest in building" them, and they "do not want others to tarnish their field by doing so" (l. 34-36).

The comparison is questionable. Indeed, historically, it is mostly German and French chemists who developed many chemical weapons (mustard gas, phosgene, etc.) during the Great War. Similarly, Zyklon B had been conceived by Walter de Heerdt, a student of Fritz Haber, recipient of Nobel Prize in Chemistry, as a pesticide. The ban on chemical and biological weapons did not spring from scientists, but from the collective consciousness, after the First World War, of the horrors of their use.

In a somehow different register, the scientific community didn't oppose, as a whole, the development and deployment of nuclear weapons. The presence of a large number of great physicists in military nuclear research centers attests to this fact.

In terms of the parallel, it is far from clear that AI will lead to autonomous weapons, and far from clear that autonomous weapons will be widely viewed as morally abhorrent, compared to the alternatives.

The Ban Claim

A Ban on Offensive Autonomous Weapons

The final paragraph proposes a *"ban on offensive autonomous weapons beyond meaningful human control"* (l. 43-44). Nonetheless, the authors should know that many discussions have already taken place, that scientists have barely participated in these discussions, and that in the United States, in 2012, the Defense Department already decided on a moratorium on the development and the use of autonomous and semi-autonomous weapons for 10 years (see above reference to the DoD Directive 3000.09). For several years, the United Nations has also been concerned about this issue. It is therefore difficult to understand the exact position of the scientific authors of the letter, especially if it does not invoke the debates that have already taken place, and to the extent that it relies on some not-altogether-germane considerations—precision, ubiquity, illicit use, firepower, etc. —such as we have explained above.

In short, the conclusion of a ban does not seem to be justified by the general argument of the letter (given the problems we have noted), nor by the novelty of the position they are staking out. There is a ban, and states are not racing ahead to deploy offensive, lethal, autonomous weapons systems. But might we be missing something? Might the authors foresee a deeper reason for scientists and technologists to eliminate the *very possibility* of an unlikely but terrifying threat?

Such would be the conclusion of an argument from the "precautionary principle," which could be the motivating principle of the ban. The precautionary principle is often invoked in environmental ethics, especially in assessing geo-engineering to combat climate change. The idea is that, while new technologies promise benefits, the threat of them going astray is so cataclysmic in terms of their costs that we must act to eliminate the threat, even when the likelihood of cataclysm is very small. The imagined threat here would be the continued development of autonomous weapon systems leading to a military AI arms race, or the mass proliferation of AI weapons in the hands of unscrupulous non-state actors, as the authors of the open letter envision.

Wallach and Allen discussed a similar argument against AI in their 2009 book *Moral Machines* [5]:

> The idea that humans should err on the side of caution is not particularly helpful in addressing speculative futuristic dangers. This idea is often formulated as the "precautionary principle" that if the consequences of an action are unknown but are judged to have some potential for major or irreversible negative consequences, then it is better to avoid that action. The difficulty with the precautionary principle lies in establishing criteria for when it should be invoked. Few people would want to sacrifice the advances in computer technology of the past fifty years because of 1950s fears of a robot takeover.

In answer to the "precautionary" challenge to autonomous weapons, it seems that Wallach and Allen provide the right balance between ethical concern and scientific responsibility:

> The social issues we have raised highlight concerns that will arise in the development of AI, but it would be hard to argue that any of these concerns leads to the conclusion that humans should stop building AI systems that make decisions or display autonomy.... We see no grounds for arresting research solely on the basis of the issues presently being raised by social critics or futurists.

Scientific Authors

Let us end by going to the beginning—with a consideration of the title (l.8-9): *"Autonomous Weapons: An Open Letter from AI & Robotics Researchers."*

Who exactly are the AI & Robotics Researchers who wrote the open letter? As a matter of fact, nothing in their presentation allows those who wrote the letter to be distinguished from those who have signed it. The question is all the more important, as some tensions within the arguments of the text suggest that some negotiations took place. In any case the open letter cannot appear as coming from *all* AI and robotics researchers. Some members of this community, both in Europe and in the United States

— not to mention the authors of this present article — have already disagreed with the content of the open letter.

To conclude, scientists and members of the artificial intelligence community may not wish to adhere to the position expressed in the open letter, not because they are interested in developing autonomous weapons, or are not "sufficiently humanitarian," but because the arguments conveyed in the letter are not sufficiently grounded in science. We think it is our duty to publicly express our disagreement because when scientists communicate in the public sphere, not as individuals, but as a scientific community as a whole, they must be sure that the state of the art of their scientific knowledge fully warrants their message. Otherwise, such public pronouncements are nothing more than expressions of one opinion among others, and may lead to more misinformation than comprehension—they may generate “more heat than light.”

It is also worth sounding another cautionary note here. When scientists decide to take the floor in the public arena, they ought to ensure that their scientific knowledge fully justifies their declarations. In these times which some commentators have declared as a “post-truth era,” the rigor of scientists' arguments is more important than ever in order to fight fake-news. This can only be ascertained after they engage in debate in their respective scientific communities, especially when some of their colleagues are not in agreement with them. Otherwise, without such open dialogue — discussions which are crucial in scientific communities to establish claims of knowledge — the public may come to doubt future declarations of scientists on ethical matters, especially if they concern technological threats. Any scientific pronouncement, whether meant for an expert community or addressed to the public, ought to take utmost care to preserve scientific credibility.

References

[1] Alexeï Grinbaum, Raja Chatila, Laurence Devillers, Jean-Gabriel Ganascia, Catherine Tessier, Max Dauchet - Ethics in robotics research: CERNA recommendations - IEEE Robotics and Automation Magazine, January 2017. DOI: 10.1109/MRA.2016.2611586

[2] Vincent Boulanin and Maaike Verbruggen – Mapping the development of autonomy in weapon systems – Stockholm International Peace Research Institute (SIPRI), November 2017. https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_0.pdf

[3] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. IEEE, 2017. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

[4] Ronald Arkin. *Governing Lethal Behavior in Autonomous Robots,* Chapman & Hall/CRC Press, 2009

[5] Wendell Wallach and Collin Allen. *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, 2009.

Appendix

Embargoed until 4PM EDT July 27 2015/5PM Buenos Aires/6AM July 28 Sydney
This open letter will be officially announced at the opening of the IJCAI 2015 conference on July 28, and we ask journalists not to write about it before then. Journalists who wish to see the press release in advance of the embargo lifting may contact Toby Walsh. Hosting, signature verification and list management are supported by FLI; for administrative questions about this letter, please contact tegmark@mit.edu.

Autonomous Weapons: An Open Letter from AI & Robotics Researchers[3]

Autonomous weapons select and engage targets without human intervention. They might include, for example, armed quadcopters that can search for and eliminate people meeting certain pre-defined criteria, but do not include cruise missiles or remotely piloted drones for which humans make all targeting decisions. Artificial Intelligence (AI) technology has reached a point where the deployment of such systems is — practically if not legally — feasible within years, not decades, and the stakes are high: autonomous weapons have been described as the third revolution in warfare, after gunpowder and nuclear arms.

Many arguments have been made for and against autonomous weapons, for example that replacing human soldiers by machines is good by reducing casualties for the owner but bad by thereby lowering the threshold for going to battle. The key question for humanity today is whether to start a global AI arms race or to prevent it from starting. If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable, and the endpoint of this technological trajectory is obvious: autonomous weapons will become the Kalashnikovs of tomorrow. Unlike nuclear weapons, they require no costly or hard-to-obtain raw materials, so they will become ubiquitous and cheap for all significant military powers to mass-produce. It will only be a matter of time until they appear on the black market and in the hands of terrorists, dictators wishing to better control their populace, warlords wishing to perpetrate ethnic cleansing, etc. Autonomous weapons are ideal for tasks such as assassinations, destabilizing nations, subduing populations and selectively killing a particular ethnic group. We therefore believe that a military AI arms race would not be beneficial for humanity. There are many ways in which AI can make battlefields safer for humans, especially civilians, without creating new tools for killing people.

Just as most chemists and biologists have no interest in building chemical or biological weapons, most AI researchers have no interest in building AI weapons — and do not want others to tarnish their field by doing so, potentially creating a major public backlash against AI that curtails its future societal benefits. Indeed, chemists and biologists have broadly supported international agreements that have successfully prohibited chemical and biological weapons, just as most physicists supported the treaties banning space-based nuclear weapons and blinding laser weapons.

In summary, we believe that AI has great potential to benefit humanity in many ways, and that the goal of the field should be to do so. Starting a military AI arms race is a bad idea, and should be prevented by a ban on offensive autonomous weapons beyond meaningful human control.

---

[3] https://futureoflife.org/open-letter-autonomous-weapons/