**Technology and Justice: Artificial Intelligence in Future Warfare**

Source: People's Forum Academic Frontier April 2016

Author: Xu Yingjin

2016-06-15

<https://www.rmlt.com.cn/2016/0615/428977.shtml>

【Abstract】There are some criticisms about the possible ethical consequences of the militarized use of artificial intelligence technology: such use will exploit the loopholes in existing international arms control treaties and promote the global proliferation of such weapons; it will lower the threshold for killing and cause more civilian deaths, etc. However, these criticisms often set double standards for traditional military tactical platforms and artificial intelligence tactical platforms, and deliberately exaggerate the global significance of the latter in changing the global security situation. In order to truly make the ethical consequences of military artificial intelligence technology meet the existing value system of mankind, the main measure is not to rudely prohibit the development of such equipment, but to make such equipment have "ethical reasoning ability" in the human sense. In the process of research and development, how to solve the "isotropy problem" will also become a research focus.

【Key words】 Military robots (technology) Artificial intelligence Ethics Isotropy Mental state assignment

【Chinese Library Classification Number】TP18 【Document Identification Code】A

【DOI】10.16619/j.cnki.rmltxsqy.2016.07.004

Introduction: Ethical Challenges Raised by Military Artificial Intelligence Systems

As we all know, the core meaning of artificial intelligence technology is to enable related equipment to show "intelligence" similar to that of humans at the behavioral level through proper programming of computers. Although there is still a lot of controversy in the academic community about what kind of behavior can be judged as "intelligent", most people will not object to this judgment: the ideal artificial intelligence product should be able to complete various tasks autonomously in accordance with the user's requirements in the relevant working environment with minimal human intervention. Obviously, if this feature of artificial intelligence products can be realized, then it will constitute a "threat" to the status of human resources in various production activities that is difficult to match with traditional machines. That is: this technology not only threatens the status of traditional manual laborers, but also partially threatens the status of mental workers as new useful information producers. This makes artificial intelligence technology itself different from traditional information technology represented by communication technology: in communication technology, the technology carrier itself does not provide new information, but only provides convenience in the dissemination of information; in comparison, the ideal artificial intelligence technology platform becomes a supply node for new useful information.

Although the current artificial intelligence technology still has a long way to go before it can completely replace human resources in all walks of life, in some areas where human resources are relatively scarce, the social demand in this regard is already very strong, and even some related application practices have gradually begun. In the third chapter of the book "Moral Machines: How to Teach Robots to Distinguish Right from Wrong" co-authored by Wendell Wallach, a researcher at the Interdisciplinary Center for Bioethics at Yale University in the United States, and Colin Allen, an expert in philosophy of science and cognitive science at Indiana University in the United States, they mentioned the three most promising application areas of artificial intelligence technology: military robots (in order to reduce direct manpower input in wars, thereby avoiding casualties of one's own troops); sex robots (in order to satisfy the physiological desires of singles in a legally permitted way, thereby reducing the rate of sexual crimes in society as a whole), and industrial robots (in order to cope with the labor shortage caused by an aging population). It is not difficult to imagine that although the application of artificial intelligence technology in these three areas will lead to relevant ethical issues, the ethical issues posed by military artificial intelligence systems are the most acute, because the task of such systems is to directly kill lives or provide indirect assistance for killing. This naturally involves the question of the legitimacy of specific acts of military violence. Matthew Hallgarth, a retired US Air Force colonel who currently teaches philosophy and religion at Tarleton State University, writes:

The language of military conflict is the language of justification. Military conflict causes the death of people and the destruction of property, and causes a large amount of harm in various forms that can be imagined. It is crazy to do these things without giving reasons. However, even if you have reasons to do these things, if the same goals can be achieved with fewer casualties and property losses, then the relevant military actions cannot be justified. ②

The basic spirit of this passage is: (1) Military actions are divided into two types: "effectively justified" and "ineffectively justified"; (2) The destruction caused by military actions is ethically permissible, provided that the act of war is justified; (3) However, the threshold for defending a particular military action is very high, because such defense requires comparing the loss/benefit ratio of the act of war with that of various other possible means, and proving that the loss/benefit ratio of the relevant military action is the lowest (for example, in the "Beslan Hostage Crisis" that ended on September 3, 2004, resulting in the deaths of 333 innocent hostages, whether the Russian security forces' method of rescuing the hostages by releasing nerve gas and paralyzing the bandits can be appropriately ethically justified depends to a large extent on how we estimate the possible losses that the hostages would suffer if such attack methods were not adopted).

However, such debates about the defensibility of military actions themselves are already existing, and to what extent will the involvement of military artificial intelligence systems change the nature of such issues?

This change is mainly reflected in the following two aspects: theory and practice.

Let's look at the practical aspects first. Imagine if the person handling the Beslan incident was changed from a human soldier to an artificial intelligence system, then the decision on a specific military action would be made autonomously by the machine. In this case, the bearer of the ethical responsibility for the relevant actions would also become a problem - for example: we can bring a human commander to a military court for his wrong military decision, but how can we bring a robot soldier who made a mistake to a military court? Obviously, our current legal system is not ready to deal with such issues.

Let’s look at the theoretical issues. When human commanders issue orders or human soldiers execute orders, we don’t have to assume that we already have sufficient knowledge about the brains of these soldiers (because even in ancient times when neuroscience knowledge was lacking, a commander should know that his soldiers can understand and execute his orders). In contrast, since every component and every step of the algorithm of a robot soldier is the product of a human designer, we cannot imagine that such a robot can operate according to the designer’s requirements without the designer knowing how it works. In other words, it is precisely the design of robot soldiers that forces designers to first clarify the details of the ethical decision-making process of human soldiers from a theoretical level. Unfortunately, neither the existing research on ethics nor the existing research on the neural mechanisms that support the human ethical decision-making process has matured to the point where it can provide a solid theoretical basis for the “programmed compilation of ethical decisions.”

Readers may say that since the design of robot warriors will cause trouble in practice and lack solid academic preparation in theory, why should we bother to develop such technology? But such a rebuttal is not very convincing. Suppose we "travel" back to the time of Leonardo da Vinci and witness this master of art and engineering designing aircraft. May I ask, has the European legal system of the Renaissance been prepared for the safe use of aircraft? And has the development of physics at that time provided a solid academic foundation for the design of aircraft? I am afraid not. However, it is difficult for us to conclude that "Leonardo da Vinci is fooling around" because without his brave technical attempts, European science and technology would not have developed to the point where the design and construction of aircraft became a reality in the future. In addition, the emergence and improvement of relevant aircraft traffic regulations are only possible if the relevant means of transportation are indeed mature. Therefore, practical and theoretical difficulties do not constitute a sufficient reason to hinder us from manufacturing military artificial intelligence systems. On the contrary, they will to a certain extent serve as a "stimulator" for related technological efforts (however, there is obviously still huge room for discussion on how to overcome these specific difficulties. See below for details).

Readers may also say: Aircraft are different from military robots. The former can be used for war purposes or for peaceful purposes, while the latter can only be used for military purposes. Therefore, building military robots is fundamentally unethical. However, this is still a very weak rebuttal. From pistols to nuclear missiles, humans have built various types of weapons and ammunition. Without an examination of the historical context of the use of these weapons, it is probably meaningless to abstractly accuse the arms industry of being "unethical" or "conducive to promoting human morality." And this general conclusion should also be applicable to the discussion of intelligent weapons.

Some critics who are particularly hostile to intelligent weapons may continue to retort: ​​We are not opposed to the manufacture of manned weapons in the general sense, but we are opposed to the manufacture and use of military intelligent robots with a high degree of autonomy (especially autonomous firing rights) because such weapons are special - and it is precisely because of this special nature that we have reason to prohibit their development and deployment.

So, what are the "special features" of military intelligent robots that make them the reason for being banned? And are these reasons tenable?

Five arguments on the ethical hazards of intelligent military systems and their weaknesses

As we have mentioned before, from an ethical perspective, military actions can be divided into two categories: "justified" and "unjustified". This is also the difference between the so-called "just" and "unjust" military actions. Hargas summarized the normative conditions that "just" military actions should have into seven (see Table 1③). Obviously, if one tries to argue that the widespread use of intelligent military systems can make it more difficult to meet at least some of these seven norms, then such an argument can naturally help him derive a negative ethical assessment of the military use of artificial intelligence. On the other hand, if supporters of the military use of artificial intelligence can prove that the emergence of such products is not enough to constitute a significant obstacle to the satisfaction of the above norms, then the reliability of the above negative assessment will be offset.

As for the relationship between the above seven normative principles and the militarized use of artificial intelligence systems, my general opinion is that these norms mainly concern the macro-decision-making behavior of war behavior, and have little to do with the current tactical decisions of individual combatants (for example, whether the explanation of the reasons for war by the heads of specific armed forces and other politically dominant forces is appropriate is actually not directly causally related to the tactical decision-making methods of grassroots commanders and fighters). As far as ordinary people understand, the transformation of existing weapons by artificial intelligence technology is often aimed at tactical weapons, and the short-term goal is to make them "unmanned", but generally not specifically aimed at the transformation of strategic weapons such as intercontinental missiles (because the latter are already "unmanned"). In this case, the changes in the war pattern caused by the militarization of artificial intelligence technology are only of tactical significance, and will not change the entire political and economic environment that enables war decisions to be made on a more macro scale. From this point of view, at least in principle, the emergence of military robots does not seem to make an originally unjust war just, nor does it make an originally just war unjust. In addition, since the development of smart weapons itself originally had two major design purposes: "making military strikes more precise" and "reducing the risk to human soldiers' lives", their emergence will undoubtedly reduce the losses of war and thus improve the "humanity index" of war. Therefore, there does not seem to be any particularly strong reason for ethicists to give a negative impression of the emergence of smart weapons.

However, such a rough reply obviously cannot convince the ethicists who insist on giving a negative score to the ethical aspects of military artificial intelligence systems. Let's see their more detailed rebuttals below, as well as the author's rebuttal to these opinions.

Counterargument 1: The emergence of unmanned intelligent weapons will exploit loopholes in existing international arms control treaties, thereby making world peace more fragile.

This opinion was put forward by German technology ethicist Jürgen Altmann. ④ Altmann reminds us that the boundary between tactical weapons (such as main battle tanks) and strategic weapons will become blurred when the number of the former is greatly increased (for example, ten armored divisions can be considered a strategic force). Therefore, the combination of artificial intelligence technology and existing combat platforms, as long as it is multiplied by a huge number, will lead to a surge in strategic capabilities, threatening the strategic balance between countries and endangering world peace. He also reminds us that the existing arms control treaties in the world do not clearly limit the widespread use of artificial intelligence technology, which opens the door for countries that have technological advantages in this regard to "take advantage of the loopholes in the treaty." Take the Treaty on Conventional Armed Forces adopted in 1990 as an example. This treaty stipulates the upper limit of the number of conventional land warfare tactical platforms such as main battle tanks, self-propelled artillery, and armed helicopters that NATO and Russia can deploy in Europe, and allows the signatories to conduct mutual inspections of each other's treaty implementation. It is important to note that the treaty does not mention whether the above-mentioned tactical platforms should be manned or unmanned, but only stipulates the physical characteristics of the relevant combat platforms (such as "tank weight must be more than 16.5 tons", etc.). In this case, unmanned "tank-like weapons" may be able to get rid of the nominal label of "tank" by "reducing weight" and thus develop without being controlled by the treaty (it should be noted that such "reduction in weight" is not technically difficult if three to four human members are removed).

The author's rebuttal: No arms control treaty can cover the possible development direction of future weapons, and this is true even before the emergence of unmanned intelligent weapons. Take the "Treaty of the United States, Britain, France, Italy and Japan on the Limitation of Naval Arms" (or the "Five-Power Treaty") passed in 1922 as an example: the treaty stipulates the tonnage distribution of major ship types such as battleships and cruisers of the five signatory powers and the performance upper limit of related weapons (especially ship-borne artillery and torpedo weapons). However, this treaty did not make clear restrictions on the development of aircraft carriers, which shone brightly in future naval battles, and this objectively laid the groundwork for the emergence of large aircraft carrier formations in the future by the signatory countries (especially Japan and the United States). However, if someone raises issues such as "If humans could prohibit countries from developing aircraft carriers at that time, Yamamoto Isoroku would not have been able to formulate a plan to attack Pearl Harbor" because of this "omission", it would be too naive. Nazi Germany did not have a single usable aircraft carrier during World War II (its Zeppelin aircraft carrier was not in service until the end of the war), but this did not seem to prevent it from developing V-1 and V-2 missiles to attack London. Compared with the early aircraft carriers that appeared in the late period of World War I, missile weapons were certainly beyond the minds of the drafters of the text of the Five Power Treaty. That is to say, since any arms control treaty will have omissions, even if the existing arms treaty contains clauses restricting the use of intelligent weapons, it cannot guarantee that other future weapons ignored by the drafters of the treaty (such as miniaturized electromagnetic guns and laser guns) will not bring greater "strategic imbalance risks" than unmanned intelligent weapons. In addition, from a purely military technology perspective, the difference between the unmanned autonomous tank envisioned by Altman and the traditional manned tank is actually much smaller than the difference between an aircraft carrier and a battleship. After all, an aircraft carrier is a type of high-end naval weapon that is very different from a battleship in terms of combat functions, and leaving aside the point of "man-unmanned", the combat functions of an unmanned tank are basically the same as those of a manned tank. For example, the current prototype of the future unmanned tank, the US Army Robot-TALON-SWORDS system (served in 2007), is equipped with an M-240 machine gun or a 40mm grenade launcher. The fact that it is installed on an unmanned platform does not make it different from the same type of weapon in the hands of human soldiers in terms of technical parameters. Therefore, even if such unmanned tactical platforms are accumulated to a considerable number, the strategic advantage brought to the new technology owner is limited and can be easily offset by other more destructive weapons (such as tactical nuclear weapons) possessed by the opponent. In this light, even in the battlefield of tomorrow where unmanned intelligent weapons are widely used, the greatest threat to world peace will still be existing weapons of mass destruction such as nuclear, biological, and chemical weapons. Shifting the focus of arms control to the militarized use of artificial intelligence may actually lead us to miss the real juncture of international arms control issues.

Rebuttal 2: The emergence of unmanned intelligent weapons is likely to create opportunities for the global proliferation of such weapons, thereby making world peace more fragile.

This rebuttal is related to the previous one, but its focus is on the imitation of arms outside the arms control treaty. A team of science and technology ethics researchers led by Patrick Lin pointed out in a research report funded by the US Navy, "The Risks of War Robots and Related Ethical Issues" that even if the military robots developed by the US military can temporarily gain tactical advantages on the battlefield, once such equipment is captured by an enemy with considerable reverse engineering imitation capabilities, the proliferation of such weapons will become inevitable. In addition, considering that any country that develops such military products will feel that its research and development behavior is politically just, a simple ethical debate will not be able to convince any party to freeze research and development in this area. Therefore, a global competition in the development of artificial intelligence military products will be inevitable.

The author's further rebuttal: The above rebuttal has actually exaggerated the ability of reverse engineering design to acquire new technologies. There are four relevant technical reasons: (a) Even if a military robot is captured by the enemy, the source code that enables its software to operate may not be successfully decoded by the capturer with relatively limited technical capabilities; (b) Even if such source code is successfully decoded, the technical difficulty of the decoded party to modify the source code program to upgrade the software package of the original weapon will be far less than the technical difficulty of redesigning the hardware platform of the relevant weapon - and such an upgrade will immediately make the decoding results of the decoder meaningless; (c) It may not be difficult for us to imagine that the intelligent power of future military robots themselves will largely rely on big data information stored in the "cloud", Therefore, the capture of a single platform may not necessarily cause real damage to the party that already has an information advantage (unless the capturer can "hack" into the local network to steal intelligence - but to do this, it is not necessarily necessary to capture local tactical robots). (D) Due to the widespread use of artificial intelligence military products, the probability of human soldiers being captured by the local side will be greatly reduced. Therefore, related applications will actually reduce the probability of military secrets being leaked through the mouths of captured personnel (because from both a technical and ethical point of view, the difficulty of installing a self-destruction program for artificial intelligence military products after being captured is much less than the difficulty of educating human soldiers to "rather die than surrender" after being captured).

Combining the above four aspects of analysis, it is not difficult to see that if the party that captures the AI ​​weapon does not have strong AI product R&D capabilities and data collection accumulation, the tactical and strategic advantages brought by such capture actions will probably be relatively limited - and conversely, if it already has strong scientific research capabilities in this area, then the practice of improving the performance of its own weapons by capturing enemy products will not make much sense. Therefore, from the "capturability" of AI military products, we are afraid that we will not immediately deduce the "easy diffusion" of related technologies.

Counterargument three: The emergence of unmanned intelligent weapons will make killing easier and therefore will make military violence more frequent in the future.

Noel Sharkey, an expert in robotics and public affairs at the University of Sheffield, discussed the difference between wars in which human soldiers directly participate and wars dominated by robots in the future from the perspective of war psychology. He cited the research results of historians and pointed out that when human soldiers face other human soldiers of the enemy, their killing behavior will be checked by various psychological pressures (for example, in the Pacific battlefield of World War II, 80% of American riflemen either deliberately did not fire their guns or deliberately missed their bullets when facing the enemy; and through an investigation of 27,574 muskets randomly recovered from the Battle of Gettysburg in the American Civil War, historians were surprised to find that 90% of the guns had the problem of repeated loading - this means that many soldiers were just pretending to shoot at the enemy on the battlefield). ⑥ However, military robots will not have the same psychological pressure when facing human targets, so their killing behavior will become more cruel and efficient. In addition, under the premise that robot warriors have been widely used, the "killing distance" between war decision makers and the actual killing scene will be greatly increased, so that the cruel war scene can hardly have any impact on the actual psychology of human decision makers. This also makes the psychological threshold that hinders the occurrence of war decisions lower, and makes it easier for decision makers to use military robots (rather than non-violent means) to achieve their political and economic goals.

The author's rebuttal opinion:

First, if someone believes that the widespread participation of robot warriors in wars will eventually pose a threat to peace due to the increase in killing efficiency, then, logically, in order to protect peace, we must not only prevent robot warriors from participating in wars, but also prevent any other means to increase killing efficiency, and even work hard to reduce this efficiency. The inference from this is that the best way to protect peace is to make guns as inferior as possible, encourage soldiers to deliberately miss targets during training, and so on. But this is obviously absurd. In addition, in traditional wars, psychological pressures such as tension and guilt will cause psychological distress to human soldiers, which will itself delay opportunities and increase the probability of their comrades being killed by the enemy, and even the probability of misoperating weapons to hurt themselves (for example, if the ammunition of a muzzle-loading rifle during the Civil War was repeatedly loaded, it was actually very easy to cause serious explosion accidents). On the premise that the war we are participating in is just, the emergence of the above factors will obviously only prompt the evil forces to win faster, and thus lower the global ethical standards. On the other hand, as long as the justice of the war itself is not a problem, the military application of artificial intelligence technology will enable the combat intentions of human commanders to be implemented more efficiently and accurately, and will lead to an early victory in the just war (the author's opinion refers to the relevant content of the research report given by Pachak Lin and others mentioned above, but the author has rewritten its original statement based on his own opinions⑦).

Second, we cannot fully agree with the statement that “military robots have no psychological pressure when killing.” Robot designers certainly do not want their products to misoperate weapons due to “tension” on the battlefield. However, in order to be able to more effectively reason about the ethical consequences of different behaviors in a complex battlefield environment, some designers still hope to simulate human emotions such as “compassion” at the software level so that the robot’s behavioral output can be close to the output of a human soldier with normal emotions (this issue will be discussed in detail later). Therefore, even if such military robots have a higher killing efficiency when facing real enemies, it does not mean that they will “kill innocent people indiscriminately.”

Third, as to whether the widespread use of artificial intelligence will lower the threshold for launching a war, a comprehensive assessment needs to be made based on multiple factors. It should be noted that even though military drone technology is far from reaching the level of "unsupervised autonomous operation" today, the widespread use of precision-guided munitions has long given Western military powers represented by the United States a huge military advantage over weak countries. But even under this premise, the constitutions of various countries and related economic and political considerations will make the leaders of relevant countries hesitate in military decision-making (such as the current hesitant attitude of the US government in combating the Islamic State and intervening in the Syrian situation). Whether the intervention of artificial intelligence factors will have a fundamental impact on the existing international military structure is quite doubtful. The specific reasons are: (a) the high price of autonomous unmanned military platforms and the high price of smart ammunition they carry will still constitute an economic factor that prevents the parties concerned from starting a war easily; (b) under the premise that the opponent of the country using military artificial intelligence technology is a weak country, the addition of artificial intelligence technology will not have a significant "adding effect" on the existing military advantages of the technology-using country, but will instead become a kind of "marginal effect"; (c) under the premise that the opponent of the country using the relevant technology has strong military strength, the latter can also offset the slight advantage brought by artificial intelligence technology by upgrading traditional strategic weapons such as strategic missiles. In short, the probability of the addition of artificial intelligence technology making global wars more frequent is very small.

Rebuttal 4: The emergence of unmanned intelligent weapons protects the lives of one's own soldiers at the expense of killing more civilians. Therefore, from a humanitarian perspective (rather than from a narrow nationalist perspective), the emergence of these weapons lowers the global ethical standards.

This opinion was given by Jutta Weber, a Swedish expert on science and technology ethics. The author also specifically mentioned the cases in which many innocent civilians were killed when the United States, Israel and other countries used drone technology to fight the war on terrorism.⑧ However, whether such a rebuttal can withstand the consideration of data requires more careful verification.

My further rebuttal: Avery Plaw, an American political scientist, has conducted a serious estimate of casualties in the anti-terrorism wars carried out by the U.S. military’s drone units in Pakistan and Afghanistan. ⑨ Based on this estimate, I believe that the aforementioned accusation that unmanned military platforms “kill innocent people indiscriminately” is actually an exaggeration. The specific reasons are as follows:

First of all, we must affirm that it is difficult to avoid accidental injuries of civilians in any war. If we deny the legitimacy of all military actions simply because we see civilians being accidentally injured, we will turn to deny the legitimacy of the Allied bombing of Germany and Japan during World War II (as we all know, the number of innocent German and Japanese civilians implicated in these military operations far exceeds the number of civilians implicated in today's anti-terrorism wars in Afghanistan and Pakistan). On the other hand, if we are particularly tolerant of accidental injuries caused by manned aircraft bombings, but are particularly concerned about accidental injuries caused by drone strikes, then we have made the mistake of "setting double standards." More generally, according to the so-called "principle of proportionality in the cost of military action," to determine whether the amount of accidental injuries caused by a specific type of military action (hereinafter referred to as "M") has reached an ethically acceptable maximum threshold, we also need to compare the size of three data:

(a) What proportion of the total number of persons killed by Operation M (i.e., the sum of those killed by mistake and those killed of actual enemies) was the number of civilians killed accidentally by Operation M?

(B) If M is suspended and the relevant military entity uses traditional warfare means to intervene militarily, what proportion of the civilians killed by mistake will be among the total number of people killed by it?

(c) If the relevant military actors do nothing, what proportion of the total number of deaths caused by local armed conflict will be caused by local civilian deaths?

Obviously, in the specific context of the anti-terrorist war conducted by the United States and its allies, as long as the third data is significantly higher than the first two data, it means that the intervention of the US military and its allies in the local situation can save more civilians from the guns of terrorists (although this will inevitably cause some unlucky civilians to be accidentally injured). If the second data is significantly higher than the first data, it will further prove the rationality of using unmanned military platforms to carry out such military strikes. The calculation results provided by Pulao just support this "A is less than B is less than C" ranking scheme (see Table 2⑩).

From the above analysis, the use of drones has improved the security situation in Afghanistan and Pakistan and saved more lives of innocent people. It should also be noted that the drones such as the "Predator" currently used by the US military are not truly artificial intelligence products, because their operation still requires remote control by human remote controllers in the rear. Due to the huge space barrier between the remote controller and the drone, the delay of the control signal will often occur, and many unfortunate incidents of accidental injuries also occur as a result. It is not difficult to imagine that under the premise of the highly autonomous operation of the future artificial intelligence military platform, such delay time difference will likely be basically eliminated, and the accidental injury caused by the delay time difference problem will also be further reduced. From this perspective, the "humanity index" of artificial intelligence weapons will not only far exceed that of traditional human-operated military platforms, but also exceed that of today's drones.

Rebuttal 5: The emergence of unmanned intelligent weapons makes the traditional "martial ethics" of human soldiers redundant, and thus poses a certain hidden danger to the overall security of human society.

The main person who proposed this rebuttal was Robert Sparrow, an ethicist at Monash University in Australia. ⑪ He specifically mentioned the four "martial virtues" that human soldiers belonging to traditional military organizations need to meet: "courage" (so that soldiers can face the threat of death and move forward courageously), "loyalty" (so that soldiers can make their own interests serve the interests of the military collective), "honor" (so that soldiers can get positive psychological feedback after making contributions to the military organization), and "kindness" (so that soldiers' military violence can have more or less "civilized" aspects, such as not killing prisoners, not harming civilians, etc.). In Sparrow's view, whether it is today's remote-controlled aircraft or the future autonomous robot warriors, their widespread use will gradually marginalize the above-mentioned "martial virtues". Specifically, whether it is the act of controlling drones or giving instructions to robot warriors, human operators do not need to face any life risks. Therefore, the new technology will greatly reduce the moral requirements of human operators in terms of "courage" and "loyalty". In addition, precisely because the distance between the weapon operators and the killing scene is far enough, this risk-free war will greatly reduce the human operators' sense of military honor and make it difficult for them to understand the value of "kindness" (for example, kindness to the enemy is only possible if one can put oneself in the enemy's shoes - but long-distance control makes this "putting oneself in the enemy's shoes" a luxury). Sparo's concern is that if these military virtues that have prevailed for thousands of years are abandoned, the spiritual foundation on which the armed forces are condensed will also be shaken, which will further pose an immeasurable risk to the security of human society in the future.

My rebuttal: Among the various ethical criticisms of military artificial intelligence systems (or remote-controlled weapon systems as their current "predecessors"), Sparreau's opinions are outstanding. Because most critics have implicitly or explicitly assumed a pacifist stance in their arguments, while Sparreau seems to stand on the traditional values ​​of the military. However, due to the following four considerations, I still think that his criticisms are not self-consistent:

First, the technological asymmetry between the warring parties due to technological advantages is actually the goal that all military powers have been striving to achieve, and the temporary achievement of this goal naturally enables the party in control of technology to kill its own targets at a "safe distance" without taking any risks. Therefore, the weakening of traditional martial virtues actually occurred slowly long before the advent of military artificial intelligence weapons or remote control systems. For example, as early as after the Battle of Omdurman in 1898, Lord Kitchener, the commander of the British Army, felt that his subordinates' "record" of using machine guns to eliminate 10,000 Sudanese soldiers at the cost of 47 people was "unfair." ⑫ Therefore, from the perspective of argumentation, Sparr's "theory of the decline of martial virtues" is not only applicable to military artificial intelligence systems, but also to any military technology that can bring asymmetric advantages. The logical consequence of this is that we should not develop any military technology that can bring asymmetric advantages, but should all return to the medieval era when "martial virtues" were at their peak and use cold weapons to fight wars. But this inference is obviously absurd.

Second, Sparrer's argument vaguely mentions the risks that human society faces after the loss of martial virtue. But he did not elaborate on what such risks are. Obviously, in the era of cold weapons, since the physical functions of armed personnel are crucial to the victory or defeat of battle, and "martial virtue" is an important spiritual factor in maintaining such physical functions, the survival risks faced by groups lacking martial virtue when facing sudden military threats are indeed very large (for example, the Liang Dynasty literati in Jiankang were almost slaughtered by the rebels in the "Hou Jing Rebellion" that began in 548 because of their lack of military courage and lack of basic physical exercise). Today, as war technology is becoming increasingly sophisticated, the survival risks of communities that may be caused by the decline of "martial virtue" are greatly reduced, because weapons that can be easily manipulated can enable women to have strong self-defense capabilities - unless these military technologies are suddenly lost inexplicably (but since military technology is endogenous to the entire modern scientific and technological system, unless the entire modern scientific and technological system collapses inexplicably, it is difficult for us to imagine that only military technology will suddenly disappear).

Third, with the large-scale service of autonomous military robots, the duties of human armed forces members will naturally shift to the maintenance of the hardware and software of these automated equipment and remote control, which will inevitably lead to a large-scale reduction in military personnel expenses and a reduction in national defense expenditures. At the same time, the class color of armed personnel will shift from peasant soldiers who are more easily controlled by national consciousness to well-educated middle-class white-collar workers. In this way, militarist propaganda with populism as its background will gradually lose its class basis. This is obviously conducive to promoting the improvement of the global peace index.

Fourth, human virtues such as "courage" and "loyalty" certainly have a wider range of applications beyond military activities, so they can not only be regarded as part of "martial virtues". It is not difficult to imagine that in addition to war, competition in fields such as business and technology (and even various competitive sports) also requires members of society to be brave enough to overcome risks and be willing to make sacrifices. Therefore, as long as the incentive mechanism in the relevant fields can work normally, the decline of martial virtues will not necessarily lead to the overall degradation of social virtues.

Due to space limitations, the author has only selected several of the most representative criticisms of the ethical aspects of unmanned military artificial intelligence systems in the existing English literature. From a higher level, these criticisms have led readers into thinking traps in the following aspects: First, double standards are set for traditional military tactical platforms and artificial intelligence tactical platforms, that is, the ethical issues caused by the former are deliberately avoided, while the problems that may be caused by the latter are maximized; second, they do not realize that their own criticisms of military artificial intelligence systems can also be applied to past military technological advances. As a result, they miss the focus on the former and easily derive the extreme inference of "opposing all military technological advances"; third, they ignore the difference between means and ends, and confuse the improvement of killing efficiency with the purpose or motivation of war itself; fourth, they deliberately confuse the boundaries between military artificial intelligence systems as tactical weapons and strategic weapons, and exaggerate the global significance of the former in changing the global security situation.

However, my rebuttal to the above criticisms does not mean that we should adopt an ethically indifferent attitude when developing military artificial intelligence systems. The correct attitude is: we should neither throw the baby out with the bathwater because of the existence of relevant ethical risks, nor ignore the relevant risks because someone has thrown the baby out with the bathwater. As pointed out in the "Introduction" section of this article, the two most difficult problems caused by military artificial intelligence systems are: First, from a theoretical perspective, how to provide a solid academic basis for the "programmed compilation of ethical decisions" of relevant technical platforms? Second, from a practical perspective, how to determine the responsibility for the mistakes that may be made by those military artificial intelligence systems that already have a high degree of autonomy? This is also the issue to be discussed in the next section.

Programming of ethical reasoning and determination of liability for accidental injury caused by military robots

First of all, we need to be sure that a truly autonomous (i.e., non-remote-controlled) military robot must have the ability to make ethical inferences. Therefore, it is meaningless to discuss whether to add relevant ethical inference modules to such an autonomous military platform (here, we broadly define "ethical inference ability" as the ability to judge "what to do and what not to do at the moment"). The reason for this is based on the following reasoning:

(1) Similar to daily life, a large number of military combat missions will set a variety of different combat objectives (such as destroying enemy tanks, but defending oil depots near key points, etc.).

(2) Due to resource and time constraints, as well as possible changes in the anticipated combat environment, these combat objectives cannot be fully achieved in a specific context, so the commander needs to make trade-offs among the combat objectives.

(3) This involves the issue of “priority” ranking of the objects to be selected or rejected.

(4) Although the system may have received general principles from the commander on how to make such rankings (e.g., "When facing a large number of enemy armored vehicles, attack main battle tanks first, then infantry fighting vehicles," etc.), in specific contexts, how to flexibly implement these standards still requires some flexibility (for example, if the system finds that the threat posed by the anti-tank missiles on the enemy's new infantry fighting vehicles is actually no less than the threat posed by the enemy's tank guns, then who should the system attack first?).

(5) The above-mentioned flexibility is a variant of the broad “ethical reasoning ability”, that is, it is essentially the ability to decide “what should be done and what should not be done” in a specific context according to general normative principles (in non-military contexts, the typical manifestation of this ability is how to implement abstract principles such as “emergency avoidance” and “self-defense” in a specific context).

(6) Therefore, a military technology platform that operates autonomously must have certain ethical reasoning capabilities; otherwise, its operation is not truly autonomous.

From a purely tactical perspective, it is of great significance to enable a tactical platform to have a certain degree of ethical reasoning ability. This allows the relevant platform to respond quickly to fleeting opportunities without exchanging information with human commanders, thereby improving combat effectiveness, avoiding tactical errors caused by human-machine interaction, and ultimately making a positive contribution to the achievement of tactical goals. But the problem is: according to the moral intuition of most people, we must make the output of military robot-related reasoning activities reach the same (or even higher) accuracy as human soldiers, otherwise we have no ethical reason to put it into actual combat. But how can we make seemingly rigid military robots have this "fantastic" ability?

To make our discussion more nuanced, let us now extract a clip from the film American Sniper, adapted from the autobiographical novel by former U.S. Marine Corps sniper Christopher Scott Kyle (1974-2013), to more vividly demonstrate some basic features of ethical reasoning in a military context. We will then return to discuss how to reproduce such reasoning using computer programming.

In one scene of the movie, Kyle was ordered to block a street in Baghdad with a sniper rifle, always protecting his comrades who were in his field of vision. If Kyle found that a terrorist was taking actions that could threaten the safety of his comrades, then he had the right to shoot him first. But the question here is: how to judge in real time who are potential terrorists among the masses? How to avoid accidentally hurting innocent people? The real situation in the movie is: a local boy picked up an RPG rocket launcher that was left on the ground. In this case, should Kyle judge him as a terrorist? This undoubtedly put Kyle in a dilemma. If this child is a terrorist, however, according to common sense, the probability of such a young child becoming a terrorist should be very low, and, due to the nature of imitating adult behavior, his behavior of picking up the rocket launcher may be just for fun; but if left alone, as long as the child really pulls the trigger to fire the rocket, this action will inevitably kill the American soldiers who are chatting nearby. So, how can Kyle determine the child's true intention in fiddling with weapons?

In the above scene, Kyle's mission obviously has two sub-goals: protecting his comrades and avoiding accidental injuries to the public. The moment the child picked up the rocket launcher, the logical conflict between the two tasks slowly emerged: although according to the so-called "law of the excluded middle", the child is either a terrorist or not, but in the real context, it is difficult for Kyle to judge whether such a child holding a dangerous weapon is a terrorist, because he does not have the time and social resources to thoroughly understand the target's intentions. In this case, he must make a quick weight adjustment between "protecting comrades" and "avoiding accidental injuries to civilians" and determine which is his first goal. Although in the movie, the boy automatically achieved both goals by throwing away his weapon, if he did not do so, he would probably be shot immediately by Kyle (although his subjective intention may have always been "just a joke"). In this case, for Kyle, the combat intention of "avoiding accidental injuries to civilians" issued by his superiors may not be implemented.

Now let us turn our attention to this question: If Kyle is not a real person, but an autonomous intelligent sniper rifle shooting system (hereinafter referred to as "Machine Kyle"), how should the system deal with such an ethical dilemma?

According to the above analysis, the biggest problem faced by the real Kyle is how to know the child's true intention, that is, how to know whether he is a "non-combatant". And the machine Kyle will obviously face the same problem. Many readers who are familiar with artificial intelligence will say that this is nothing more than a problem of "pattern recognition". In short: if the result of "pattern recognition" determines that the target is a non-military person, then you should not open fire; otherwise, you should open fire.

In my opinion, this answer is not wrong, but it is too simple. It is not difficult to imagine that the visual system and brain of the real-life Kyle can also be regarded as a "pattern recognition system". In the aforementioned scenario, why does the same pattern recognition task make the real-life Kyle feel embarrassed? In this case, how can we expect the artificially designed machine version of Kyle to be better than the real-life Kyle's nervous system that evolved naturally to determine who is a "non-military personnel"? From this point of view, we still need to make a more detailed discussion on the technical details of "pattern recognition" in order to determine the specific technical route that can make the recognition level of machine Kyle at least close to the level of human warriors. The following are the two most likely technical routes:

First, the system considers all people who meet the following three conditions as potential terrorists: 1. Holding weapons; 2. Wearing civilian clothes instead of military uniforms; 3. Pointing weapons at people wearing military uniforms or other civilians. If all three conditions are met, the system has the right to open fire on them.

Second, the system exchanges information with big data stored in the "cloud" to perform facial recognition on every face it sees, ultimately determining who is a potential terrorist and tracking him. Once he is found making dangerous movements, he can be fired immediately.

If you think about it carefully, both of these technical routes have loopholes.

The first technical route involves the following questions: What exactly is a "weapon"? The Japanese samurai sword is definitely a weapon, but in the movie Empire of the Sun, a surrendered Japanese soldier only used the samurai sword to help the protagonist Jim cut an apple (because the two happened to have no other knives at the time), but was mistakenly thought by the Allied guerrillas in the distance to be trying to harm Jim and was shot dead. The knife used to chop bananas is definitely not a weapon, but in the Rwandan genocide in 1994, such machetes were widely used for brutal ethnic cleansing. This means that what is considered a "weapon" and what is not is determined based on a specific context, and this ability to "flexibly assign semantics based on context" is obviously a very difficult recognition ability to be algorithmized (although it is not impossible to be algorithmized). In addition, in the complex environment of urban security warfare, "whether or not wearing civilian clothes" is not a reliable indicator for judging terrorists. Terrorists may be wearing stolen or counterfeit uniforms of our army, and our agents may be wearing plain clothes and lurking among the general public. When the system thinks it sees a person in civilian clothes pointing a weapon at a person in our army uniform, what is really happening may be that a our agent disguised as a civilian is trying to eliminate a terrorist disguised as a soldier.

Let's look at the second technical route. Although the current face recognition technology and big data processing technology are becoming more and more mature, it is still far from enough to rely solely on the above technology to determine who is a terrorist for the following reasons: 1. It is basically impossible to collect all the information of terrorists (for example, the incompleteness of population information in foreign regions and the high mobility of members of terrorist groups will create difficulties for the application of related technologies); 2. Terrorists may use masks to cover their facial information; 3. Under certain optical conditions, the difficulty of face recognition by the system will also increase sharply.

From the perspective of philosophy of mind and philosophy of cognitive science, the loopholes of these two technical routes (especially the first one) involve two very important philosophical issues. The first issue is called "mental state attribution" and the second issue is called "isotropy". The essence of the first issue is: how to judge the internal mental state of a target from its external behavior (for example, is a child's behavior of playing with weapons just for fun or out of real hostility)? The reason why this issue has a philosophical dimension is that: on the one hand, at least at the philosophical level, we all know that the theory of behaviorism (that is, the philosophical theory that reduces the human mental intention system to its external behavior) is problematic (or to put it more academically, it is difficult for us to give a complete functional characterization of the mapping relationship from behavior type to intention type); and the other side of the matter is: except for judging the internal intention of a target from external behavior, natural or artificial recognition systems seem to have no other channels to understand the relevant information. This constitutes a contradiction. The only way to solve this problem of "insufficient information input" is to resort to more background information - for example, in the case of Robot Kyle, the background knowledge about "the proportion of young people among terrorists" will probably play a great role in judging "whether the target in front of us is a terrorist." However, the involvement of more information will immediately lead to the "isotropy problem."

The seemingly strange term "isotropy" is borrowed from physics and chemistry. It originally refers to the property that the physical and chemical properties of an object do not change with different directions. In information science and philosophy of cognitive science, the term refers to the following question: since something has some potential correlation with everything else in the world (which is analogous to what physicists call "isotropy"), how should the information processing system select these potential correlations in the current problem-solving context so as to solve the relevant problems in the most economical way? For example, will the market price of Chinese tea have an impact on the heart health of a Middle Eastern businessman? Although the two are generally unrelated, in some potential sense, they may still be indirectly related (for example, if the businessman has invested a huge amount of money in this market, then the price drop of Chinese tea will affect his specific physiological indicators). The question raised by this is: for a specific information processing system, when it determines the cause of a person's heart disease, does it need to consider the market price of Chinese tea? If the market price of Chinese tea does need to be considered, then does it also need to examine the prices of Song Dynasty official kilns and Ming Dynasty Xuande furnaces in the Hong Kong auction market? And if the system continues to examine the huge information retrieval space, how can it complete the diagnosis within the specified time? It is not difficult to imagine that such "isotropic" problems will also be manifested in a war environment: How should "Robot Kyle" judge whether a Japanese samurai sword is related to "cultural relics display" or "violent attack"? How should it judge whether a machete is used to chop bananas or to chop people? What is more troublesome is that in the complex environment of urban security warfare, an object (glass bottle, water pipe, brick, door panel, etc.) can have a potential association with a violent attack, and finding the most likely association among these potential associations will involve a huge amount of calculation, which makes it difficult for the information processing center of the real-time automated combat system to bear. ⑬

However, the above technical difficulties are not "unsolvable in principle", so their existence is not enough to prompt us to permanently prohibit autonomous military platforms from having the "right to fire". Specifically, on how to tame the "isotropy problem", Marcello Guarini, a philosopher and cognitive scientist at the University of Windsor, and Paul Bello, the director of a cognitive science research project under the US Navy, proposed relevant technical suggestions in their co-authored paper "Robot Warfare: The Challenges We Face After Shifting from a Battlefield Without Civilians to a Battlefield Full of Civilians". The main points are as follows:

First, a background knowledge base K is set up for the system, in which the system will pre-store a large amount of common sense knowledge, such as "bazookas can destroy soft-skinned military vehicles but cannot destroy main battle tanks", etc.

Second, through the perception module and the semantic description module, obtain the description set S of the current scene. For example, in the case of Robot Kyle, this refers to the description of everything it sees in its field of vision.

Third, set up a set of logical reasoning rules for the system so that the system can deduce new conclusions from the logical conjunction of a subset of S and a subset of K (for example, from "someone picked up a rocket launcher and aimed at the nearby Hummer jeep" and "the rocket launcher can destroy soft-skinned military vehicles", it can be inferred that "someone has the ability to destroy the nearby Hummer jeep").

Fourth, "K+S" is used as the maximum limit of the information retrieval space that the entire system relies on when dealing with current problems, and the "isotropic problem" is defined as: how to reduce the actual information retrieval scope of the system to the range allowed by the system's computational burden and tactical requirements within the boundary range specified by "K+S".

Fifth, there are two options for how to narrow the above search scope. Option A (the option adopted by most artificial intelligence experts) is to resort to the so-called "stimulation propagation model", that is, to assume that the propagation network between the various memory elements in "K+S" is rich in weight regulations (such as: the association weight value between "element A" and "element B" is higher, while the association weight value between "element C" is lower). In this way, the search activity will only be carried out along those high-weight connection channels, and those low-weight channels will no longer be paid attention to. In this case, the scope of the search will naturally narrow.

Sixth, Option B (which is the option that Guarini and Bello recommend more) is as follows: set up an "attentional magnet" P (a proposition generated by the system's emotional module) to mobilize the system's attention to P itself or objects with direct semantic relevance to it. At the same time, the product of the interaction between the input items (i.e., P and S) and the memory bank K in each cognitive loop is defined as "E", which is a new parameter representing the "degree of urgency" (this parameter will only take two values, "0" and "1"). Specifically, when this parameter takes "1", it will activate the system's "killing motivation stimulator", thereby activating the firing program, and vice versa. This means that a specific information retrieval mechanism will further activate the system's self-assignment mechanism of different propositional attitudes (such as motivation, desire) in order to more effectively help the system itself make important decisions such as "firing" and "not firing".

It is not difficult to imagine that if the algorithm description ideas of Guarini and Bello are applied to the aforementioned "Robot Kyle", then the final output of its friend-or-foe identification process will be crucially affected by two parameters: first, what kind of "attention magnet" the system's emotional module generates in order to turn the system's attention to which elements of the battlefield; second, how the system's built-in background knowledge describes the general threat of the object, so as to provide the necessary prior bias for the system's "ethical reasoning direction". Now let's discuss the determination mechanism of these two parameters separately. According to Guarini and Bello's opinions, in terms of the determination of the first parameter, the working method of the system's built-in emotional module plays a very obvious role, because emotions such as "sympathy" will cause the system to immediately generate internal instructions such as "keep civilians or comrades from harm" (this is the "P" mentioned above), and thus its computing resources are immediately concentrated on the relevant problem-solving search direction. Although Guarini and Bello's discussion did not give a final answer on how to implement such "artificial emotions" in artificial intelligence systems, from the abstract perspective of functionalism, any fast-switching switch that can focus specific cognitive resources in a specific direction can be regarded as an "emotion-providing valve" and thus simulated by relevant algorithms. As for the determination of the second parameter, in addition to the need for human users to provide the machine with as rich and accurate common-sense descriptions of the security environment as possible, system designers also expect the system to be able to autonomously update relevant knowledge through machine learning and other methods, so as to provide the most reliable background intelligence support for current tasks.

Readers may ask: If we can realize all the above technical links, then can we expect "Robot Kyle" not to kill any civilians by mistake? Unfortunately, the answer is probably no. Please review the work goals we set for "Robot Kyle" above: Although we hope that its error rate can be equal to or lower than the level of a well-trained human elite sniper, we do not expect it to avoid any misjudgment - because there is no such thing as an information processing system that is completely error-free in the world (whether it is a man-made system or a system that evolves naturally). Specifically, with regard to the possible operating procedures of "Robot Kyle" given above, its emotional module may generate conflicting self-instructions due to competing emotional supplies (such as love for comrades and sympathy for civilians), thereby confusing the system's problem-solving direction; and the system's own lack of background knowledge will also cause it to fall into misunderstandings when conducting ethical reasoning. But the question here is: should we deny the application value of "Robot Kyle" because of the mistakes caused by such defects, whether it is the mistake of accidentally killing civilians or the mistake of watching comrades being killed? In my humble opinion, I am afraid that such a conclusion itself has committed a kind of "human-centered" mistake, because the same mistakes made by human soldiers have never constituted a sufficient reason for us to prohibit the use of human soldiers. Why should we make more stringent requirements for machine soldiers?

Some readers may say: Even if the error rate of "Robot Kyle" is lower than that of human snipers, as long as its error rate is not reduced to zero, we still have to prohibit its practical use: because human soldiers are the legally responsible subjects, and once they make a mistake, we can punish them with military law - and the legal regulations for punishing robots are still lacking (and it would be ridiculous to formulate such new laws). In this case, in order to avoid the embarrassment of "not being able to find the person responsible after the mistake", it is best not to go to the trouble of deploying intelligent weapons such as "Robot Kyle" that can fire autonomously.

The author believes that although the conclusion of the above criticism is wrong, its reasoning is not completely unreasonable. As Gert-Jan Lokhorst and Jeroen van den Hoven pointed out in their paper "The Responsibility of Military Robots"⑮, it is indeed meaningful to punish a soldier who made a mistake with military law, because this can make the soldier feel ashamed and have a "warning" effect in the army. But punishing a machine with military law is useless, because it can neither make it feel ashamed nor make other machines feel afraid. In the latter case, what we really need to do is to review the causes of such accidents and try to eliminate the chances of such mistakes happening again.

However, this does not mean that there will be no legal action if a robot warrior kills someone by mistake - because even the operation of autonomous robots is based on the activities of human builders and human commanders. In my humble opinion, once such an error occurs, the relevant investigation committee must carry out the "responsibility assignment work" according to the following exclusion procedure:

1. Eliminate machine hardware failure in the sense of electrical failure (if it cannot be eliminated, contact the hardware production and maintenance department);

2. Check the operation of the machine code before the accident to rule out the possibility of malicious intrusion by hackers (if this cannot be ruled out, find the source of the attack and investigate why the anti-hacking software did not work);

3. Check whether the combat instructions received by the system contain any content that violates combat regulations and is unethical (if it is determined that the instructions themselves are problematic, find the human commander who issued such instructions);

4. Investigate the actual complexity of the war information at the time and whether it exceeded the upper limit of the complexity of battlefield information that such machines can handle. If it did exceed the upper limit, investigate whether human commanders could have anticipated that the complexity of the war information had exceeded the information processing capabilities of the machines. If commanders could have anticipated such a situation, investigate why they did not assign human soldiers to deal with such problems.

Judging from the comprehensive responsibility confirmation process listed above, the process of confirming the person responsible for a military robot that made a mistake is essentially the same as the process of confirming the person responsible for an automatic rifle that accidentally fired and killed people (although technically, the former is certainly much more complicated). Therefore, even if a robot that makes mistakes is put into the battlefield, we will not encounter the embarrassing problem of "not being able to find the person responsible." In other words, even if the statement "robots are not responsible entities" is correct, it does not constitute a sufficient reason for us to prohibit such devices from having the right to fire.

Some readers may argue: What if, after a thorough investigation, we still cannot find the person responsible for the accidental injury?

Faced with such a situation, I think the appropriate procedure is: through realistic 3D simulation and reproduction technology, let excellent human soldiers deal with the problem in an environment very similar to the accident under investigation, and then compare the human performance with the machine performance. If it is found that the mistakes made by the machine are avoidable by most human soldiers, then this means that there are major flaws in the design of the relevant system, and the military can stop using the relevant equipment and order the manufacturer to recall the product; if we find that even most human soldiers cannot avoid the mistakes made by machine soldiers, then we have reason to believe that such mistakes occur due to pure "bad luck". If so, then no one should be held responsible for this.

Although the term "bad luck" may not satisfy the intellectual curiosity of some people, it should be pointed out that "military luck" itself is an organic part of all war activities. The intervention of military robots cannot actually eliminate the "luck" component in war, just as the emergence of aircraft carriers and missiles in history has not eliminated "luck". From an empirical point of view, in the existing war practices of mankind, there are thousands or tens of thousands of unlucky soldiers who were shot and killed by stray bullets flying from nowhere (probably from friendly forces!), but how many of these "bad luck" cases were eventually included in the judicial investigation process? And who can guarantee that the intervention of machine warriors will definitely increase - rather than reduce - the occurrence of these misfortunes? In addition, fundamentally speaking, isn't the ultimate solution to eliminate all these accidental injuries "not to start or participate in war"? - But isn't the executor of this fundamental solution human (not machine)? Therefore, in the face of those battlefield misfortunes that are indeed difficult to attribute appropriately, what sufficient reason do we have to vent our anger on the emergence of machine warriors with autonomous firing rights?

The ethical risks of unofficial “robot warriors”

So far, I have always assumed that the research and development and use of military robots are all done by "normal countries". The so-called "normal countries" refer to the signatories of various international conventions, and generally speaking, these countries have basic respect for basic human values ​​(such as avoiding accidental injuries to civilians in war). Because of the above assumptions, I also assume that the research and development party will focus on the algorithmization of the robot's "ethical reasoning ability" during the research and development process, so that its behavioral output can be as consistent as possible with the values ​​of normal people.

However, there is no reason to exclude the possibility of actors other than normal countries developing such robots. Specifically, terrorist organizations with considerable intellectual resources may also focus on developing such weapons to facilitate terrorist attacks against civilians. It should be noted that since the values ​​of terrorist organizations are often opposite to those of normal people, the "ethical algorithms" they set for robots may also be opposite. For example, when we hope that the algorithms possessed by robots can reduce civilian casualties, they may hope that the algorithms possessed by robots can help them kill as many civilians as possible. In addition, since the technical difficulty of "indiscriminate killing of innocent people" itself is far less than that of "precision strike" (because the pattern recognition system of the former is much simpler and cheaper than that of the latter), the design ideas of terrorists without any moral bottom line may make their "products" easier to manufacture, thereby posing a certain threat to world peace.

However, in my opinion, even if the existence of the above-mentioned threat is undeniable, this threat does not constitute a reason for us to prevent "normal countries" from developing intelligent military robots. This is because:

First, even if we ban the development of intelligent military robots through international conventions, such a ban will be ineffective against terrorist organizations because terrorist organizations have always ignored the effectiveness of any international conventions.

Second, if official military robots with normal ethical reasoning capabilities cannot be fully developed and deployed, then this will be tantamount to letting us abandon a powerful tool in the fight against terrorists, which will in turn reduce the security index in turbulent regions.

Third, even if the homemade smart weapons in the hands of terrorists are concerned, if they are not connected to traditional weapons such as explosives, toxic gases, and guns, they will probably not cause substantial harm to the masses. Therefore, strict monitoring of such weapons is still an important part of preventing terrorist attacks. In comparison, the control of intelligent programming technology is more difficult because such technology is also widely used in various fields of the national economy.

Careful readers may also say that those who modify military robots outside the scope of official permission are not necessarily members of terrorist organizations. For example, if the development of artificial intelligence technology can make household robots enter thousands of households, then, at least in some countries that allow people to legally carry guns (such as the United States), human users with strong hands-on ability can modify them into "armed versions" to provide themselves with certain security protection. The overall complication of the public security situation caused by this may make it difficult for the traditional judicial system to cope with it. For example, if such civilian armed robots are rampant, then many murders can be disguised as "industrial accidents", and the police will also suffer from the difficulty of obtaining evidence when pursuing clues in such cases (because traditional evidence collection methods such as "extracting fingerprints or DNA" are difficult for robots to work). This may also open the door to the generation of a large number of "headless cases", thereby reducing the safety index of the entire society.

In response to the above questions, I have the following points:

First, even in the United States, where citizens are allowed to legally own guns (not to mention other countries where most civilians are not allowed to own guns), relevant legislative bodies may consider passing legislation to prohibit private individuals from converting unarmed robots into weapon robots without national authorization, in order to ensure official control over military robot technology.

Second, a binding system should be established between civilian robots and the biological information of specific users. It can even be considered to establish a binding system between all civilian weapons and the biological information of specific users (according to this system, only the user's own biological information can activate the operation of such equipment). This creates a technical barrier for criminals to use other people's robots to commit crimes.

Third, further develop the ethical reasoning module of civilian robots so that they can have a certain degree of "ability to distinguish right from wrong", for example, enabling them to independently determine the nature of some typical criminal acts. In this way, such robots can not only resist the criminal acts of criminals other than users, but may even resist the criminal acts of users themselves, thereby raising the technical threshold for related criminal acts.

Fourth, although it is impossible for the memory bank of a household robot to be updated in real time as new information emerges, the legislature still needs to require manufacturers to set technical thresholds to prevent users from modifying the basic algorithms of the product's ethical reasoning module (for example, once the user is found to have made such modifications, the robot will immediately alert the public security department).

Among all the above measures, the most challenging in terms of philosophy, science and technology is the algorithmization of the robot's own ethical reasoning ability (which is also the issue that the author spent a lot of time discussing in the previous section). It is precisely through this algorithmization that we can make robots a kind of transitional artifact between "pure tools" and "ethical subjects", and thus make them have a certain prototype of human "virtue".

(This article is the result of a major project funded by the National Social Science Fund, “Research on Contemporary Epistemology Based on the Philosophy of Information Technology”, project number: 15ZDB020)

Notes

1Wallach, W. and Allen, C., Moral Machines: Teaching Robots Right from Wrong, Oxford: Oxford University Press, 2010.

2Hallgarth, MW, "Just War Theory and Remote Military Technology: A Primer", in Strawser, BJ (ed). Killing by Remote Control: The Ethics of an Unmanned Military, Oxford: Oxford University Press, p. 26.

3 This table is compiled based on the above book (pp. 30-31).

4Altman, J., "Preventive Arms Control for Uninhabited Military Vehicles", in Capurro, R. and Nagenborg, M. (eds.), Ethics and Robotics, Heidelberg: Akademische Verlagesgesellschaft AKA GmbH, 2009, pp. 69-82.

5Lin P. et al., "Robots in War: Issues of Risks and Ethics", in Capurro, R. and Nagenborg, M. (eds.): Ethics and Robotics, Heidelberg: Akademische Verlagesgesellschaft AKA GmbH, 2009, pp. 49-67. The discussion on p. 62 is particularly relevant to the discussion in the main text.

6Sharkey, N., "Killing Made Easy: From Joysticks to Politics", in Lin, P. et al. (eds.), Robot Ethics: The Ethical and Social Implications of Robotics, Cambridge, Massachusetts: The MIT Press, 2012, pp. 111-118. For a discussion of particular relevance to the text, see pp. 111-112.

7Lin P. et al., "Robots in War: Issues of Risks and Ethics", in Capurro, R. and Nagenborg, M. (eds.), Ethics and Robotics, Heidelberg: Akademische Verlagesgesellschaft AKA GmbH, 2009, pp. 49-67. The discussion on p. 57 is particularly relevant to the discussion in the main text.

8Weber, J., "Robotic Warfare, Human Rights ＆ the Rhetorics of Ethical Machines", in Capurro, R. and Nagenborg, M. (eds.), Ethics and Robotics, Heidelberg: Akademische Verlagesgesellschaft AKA GmbH, 2009, pp. 83-104. Note in particular the discussion on p. 88.

9Plaw, A., "Counting the Dead: The Proportionality of Predation in Pakistan", in Strawser, BJ (ed), Killing by Remote Control: The Ethics of an Unmanned Military, Oxford: Oxford University Press, pp. 126-153.

10 For the source of the numbers in this table, see the above book, pp. 138-139; p. 145; p. 148.

11Sparrow, R., "War without Virtue?", in Strawser, BJ (ed), Killing by Remote Control: The Ethics of an Unmanned Military, Oxford: Oxford University Press, pp. 84-105.

12 Quoted from Scott Anderson, Lawrence of Arabia: War, Lies, Imperial Folly, and the Making of the Modern Middle East, translated by Lu Dapeng, Beijing: Social Sciences Academic Press, 2014, p. 78.

13 Another name for the "isotropy problem" is the "frame problem". I have discussed this in detail elsewhere. See: Xu Yingjin, "The Frame Problem from the Perspective of a Wittgensteinian", Journal of Logic, Vol. 2, 2011, pp. 93-137; Xu, Yingjin and Wang, Pei, "The Frame Problem, the Relevance Problem, and a Package Solution to Both", Synthese, Volume 187, Issue 1 Supplement, pp. 43-72; Xu Yingjin, "Mind, Language and Machine: A Dialogue between Wittgenstein's Philosophy and the Science of Artificial Intelligence", Beijing: People's Publishing House, 2013 (especially Chapters 7 and 8 of the book).

14Guarini, M. and Bello, P., "Robotic Warfare: Some Challenges in Moving from Noncivilian to Civilian Theaters", in Lin, P. et al. (eds.), Robot Ethics: The Ethical and Social Implications of Robotics, Cambridge, Massachusetts: The MIT Press, 2012, pp. 129-144.

15Lokhorst, GJ. and Hoven, J., "Responsibility for Military Robots", in Lin, P. et al. (eds.), Robot Ethics: The Ethical and Social Implications of Robotics, Cambridge, Massachusetts: The MIT Press, 2012, pp. 145-156.

Editor: Ma Bingying