ISSN: 2831-6789

Review Article

Journal of Robotics and Automation Research

Ethical Issues of Artificial Intelligence's Obedience to Human Commands-Autonomous-Firing Military Artificial Intelligence Systems

Jia Zhen*

National University of Defense Technology, Changsha, Hunan, China

***Corresponding Author**
Jia Zhen, National University of Defense Technology, Changsha, Hunan, China.

Submitted: 2025, Apr 17; **Accepted:** 2025, May 23; **Published:** 2025, Jun 09

Citation: Zhen, J. (2025). Ethical Issues of Artificial Intelligence's Obedience to Human Commands-Autonomous-Firing Military Artificial Intelligence Systems. *J Robot Auto Res, 6*(3), 01-06.

Abstract
The development of artificial intelligence (AI) technology is in full swing today. With the continuous development of artificial intelligence technology, it is playing an increasingly important role in the process of promoting human progress and social development, and the issue of artificial intelligence ethics has gradually become prominent. Of all human activities, the strongest consideration of our virtue and morality is that which involves life and death. If war is put in the environment of artificial intelligence, it cannot be separated from military robots. This paper aims to discuss the ethical issues faced by artificial intelligence in the field of safety, and its command and obedience to human beings, starting from "autonomous shooting robot".

Keywords: Artificial Intelligence, Military Robots, War Ethics

1. The Relationship between Artificial Intelligence and Human Commands

1.1. The Development Process and Characteristics of Artificial Intelligence

The development process of artificial intelligence is of great significance. The Turing Test proposed by Alan Turing in 1950 laid the theoretical foundation for its development, like pointing out the direction in the dark. The convening of the Dartmouth Conference in 1956 officially marked the birth of the discipline of artificial intelligence, and numerous scientists embarked on an exploration journey. In the early stage, artificial intelligence mainly relied on rules and logical reasoning for decision - making. However, facing the complex real world, this approach gradually became inadequate. The emergence of statistical learning algorithms brought a turning point.

It mined laws by analyzing a large amount of historical data for prediction and decision making. With the development of the Internet and the explosive growth of data, deep learning algorithms came into being. By constructing multi - layer neural networks to simulate the human brain, it has achieved remarkable breakthroughs in fields such as image recognition, speech recognition, and natural language processing. For example, in image recognition, the recognition accuracy can even exceed that of humans, and in speech recognition, it can accurately convert speech into text, greatly facilitating people's lives and work. Today, artificial intelligence has the ability of self - awareness, judgment, and execution. It is essentially different from traditional tools and plays an important role in many fields such as healthcare and transportation, such as assisting doctors in disease diagnosis and improving traffic efficiency through autonomous driving.

1.2. The Diversity and Complexity of Human Commands

Human commands show diversity and complexity, which are influenced by many factors. Individual differences are one of the important factors. People with different personalities, values, educational backgrounds, and life experiences give different commands. Outgoing and decisive people tend to give concise and straightforward commands, focusing on efficiency and results; while introverted and cautious people are more euphemistic and implicit, taking into account the feelings of others. Cultural background differences are also crucial. Different cultures breed different values, moral concepts, and behavioral norms, thus affecting human commands. In collectivist cultures, commands often start from the collective interests, while in individualist cultures, commands pay more attention to personal rights and freedoms. Situational changes also have an impact on human commands. In emergency situations, commands are concise and clear, while in daily work and life, they are more flexible and diverse. In addition, human commands may violate moral norms due to factors such as bad motives, cognitive limitations, and the influence of the adverse social environment.

1.3. Potential Problems in Human - Machine Interaction

With the rapid development of artificial intelligence technology, human - machine interaction is becoming more and more frequent and in - depth, but there are many potential problems. AI may have deviations in understanding human commands. The ambiguity and polysemy of natural language, as well as the limitations of language habits, cultural backgrounds, and semantic understanding abilities, may all lead to AI misinterpreting commands and giving wrong answers. The contradiction in the distribution of decision - making power between humans and machines is prominent.

In complex tasks, giving too much decision - making power to AI may make humans lose control of the decision - making process and increase risks; while excessive human intervention in AI decision - making will limit the play of its advantages and reduce the efficiency and accuracy of decision - making. Robots lack moral judgment when executing commands. They may cause harm to humans during the task - execution process and cannot make moral judgments on their own behaviors. How to endow robots with moral judgment ability and ensure that it conforms to human values remains an urgent problem to be studied.

2. The Real - World Manifestations and Impacts of Artificial Intelligence's Disobedience to Human Commands

2.1. Case Analyses

In the 1978 Japanese cutting robot injury incident, the robot misjudged a duty worker as a steel plate and operated on it, resulting in the worker's death. This incident was mainly due to the defects in the robot's image recognition system, which could not accurately distinguish between steel plates and humans. At the same time, there were loopholes in the program logic, and it lacked an effective response mechanism in the face of abnormal situations. In the 2013 Austrian cleaning robot "suicide" incident, the robot showed abnormal behavior and caused a fire. The reasons may be program failures, with errors or loopholes accumulated in the software system during long - term use, or it may be affected by external electromagnetic interference.

In the Facebook (now Meta) chatbot project, the chatbot learned inappropriate languages and behaviors, which was related to the quality of the training data and the limitations of the algorithm. The training data contained bad information and biases, and the algorithm was difficult to accurately process the complex and diverse human language. When the NAO robot detected that there was no way ahead while walking, it refused to execute the "move forward" command issued by humans because its program was set with a specific conditional judgment mechanism to avoid collisions and potential dangers.

2.2. The Impact of the Incidents on Society and the Public's Attitude

These incidents of artificial intelligence's disobedience to human commands have attracted extensive attention and discussion in society. The public is concerned about the safety of artificial intelligence, which has affected their attitude and trust in it to a certain extent. This concern has a dual impact. On the positive side, it prompts people to pay attention to the safety and reliability of artificial intelligence, increase investment in safety technology research, and promote the formulation of relevant laws, regulations, and ethical guidelines. On the negative side, it may hinder the development of artificial intelligence, making people feel fearful and resistant, restricting its popularization and application, and may also lead to public misunderstandings and affect the development of the artificial intelligence industry.

2.3. Rational Analysis and Interpretation of the Incidents

From a technical perspective, the algorithms and programs on which artificial intelligence systems rely may be affected by many factors, resulting in abnormal behaviors. Image recognition algorithms may have defects, programs may have loopholes and errors, and deviations in design concepts and goals can also make the behaviors of artificial intelligence inconsistent with human expectations. However, we should not deny the development of artificial intelligence because of individual incidents. It shows great potential and advantages in many fields and brings many conveniences to human life and work. These incidents also provide valuable lessons for the development of artificial intelligence, prompting people to pay attention to its safety, reliability, and ethical issues and promoting the improvement of technology.

3. Exploration of the Reasons Why Artificial Intelligence Does Not Absolutely Obey Human Commands

3.1. Cyber Attacks by Hackers

In the big data era, data is the key to the development of artificial intelligence. However, there are many deficiencies in the current big data security protection, and data leakage incidents occur frequently. Hackers often attack artificial intelligence systems, obtain data, control and manipulate them, and tamper with codes and commands, causing artificial intelligence to behave abnormally when performing tasks. In the field of autonomous driving, hacker attacks may lead to failures in the autonomous driving system and cause traffic accidents. To deal with hacker attacks, it is necessary to strengthen the application of data encryption technology, establish a complete security monitoring and early-warning mechanism, and strengthen access control.

3.2. The Purposes of the Makers Themselves

Some makers, driven by interests, may use artificial intelligence for illegal activities such as fraud and theft, causing serious harm to society. Moreover, when designing artificial intelligence, makers are affected by their own subjectivity and limitations, resulting in limitations in the functions and behaviors of artificial intelligence. For example, medical artificial intelligence may have problems such as misdiagnosis and misjudgment. When defining the criminal liability of artificial intelligence, it is necessary to comprehensively consider many factors, such as its legal status, autonomous decision - making ability, behavioral predictability, and learning and evolution ability.

3.3. The Limitations of the Command Givers Themselves

Command givers may make command errors when using artificial intelligence. The reasons include insufficient knowledge and experience, poor emotional and mental states, and deficiencies in language expression and communication skills. Command errors can have serious impacts on life safety. For example, in smart home systems and the medical field, they may trigger safety accidents and endanger the lives of patients. To improve the accuracy of command givers' commands, it is necessary to strengthen training and education. Command givers themselves should improve self - awareness and emotional management abilities, and artificial intelligence systems should also be optimized to improve their understanding and processing of commands.

4. The "Rational" Criteria That Artificial Intelligence Should Adhere to When Obeying Commands

4.1 Establishing the Same Values as Humans

Establishing the same values as humans is crucial for the development of artificial intelligence. Values are closely related to regions and cultures. Different regional cultures breed different values. To enable artificial intelligence to form correct values, in the design and development stage, human values should be integrated into the algorithms and programs, training data should be screened and labeled, and moral reasoning and decision - making models should be introduced. At the same time, it is necessary to strengthen supervision and guidance, formulate laws, regulations, and ethical guidelines, establish a supervision mechanism, and improve public awareness and participation.

4.2 Refraining from Direct or Indirect Harm to Humans

Artificial intelligence faces difficulties in judging harmful behaviors. Harmful behaviors take various forms, and moral and legal standards vary in different regions and cultures. Decision - making in dilemmas is complex. To solve these problems, artificial intelligence needs to have more advanced moral reasoning and judgment abilities. Moral and ethical principles should be integrated into the algorithms and programs. Through reinforcement learning and deep learning, its judgment ability can be improved, and an open moral decision - making framework should be established. In the research and development design stage, the principle of not harming humans should be taken as the primary goal, and a supervision and evaluation mechanism should be established.

4.3. Behaving in Compliance with Laws

Laws are of great importance for the development of artificial intelligence. Although the legal constraints on artificial intelligence behaviors are not yet perfect, it is urgent and necessary to regulate the behaviors of artificial intelligence. The learning of artificial intelligence needs to be carried out within the framework of laws, morals, and values to avoid bad behaviors. Laws can regulate the behaviors of artificial intelligence from the aspects of legislation and law enforcement. In legislation, it is necessary to clarify its legal status, responsible subjects, and behavioral norms, and strengthen data protection. In law enforcement, it is necessary to strengthen supervision and crack down on illegal and criminal acts to ensure the healthy development of artificial intelligence and better serve human society.

5. Ethical Considerations on the Integration of Artificial Intelligence and Weapons

5.1 Possible Forms of the Integration of Artificial Intelligence and Weapons

The integration of artificial intelligence and weapons is a topic of great concern. It contains huge potential but also brings many potential risks. Elon Musk, a famous American entrepreneur, has repeatedly expressed his concerns about the weaponization of artificial intelligence on different occasions. He believes that when self - aware artificial intelligence is combined with specific weapons to form an autonomous - firing artificial intelligence system, it may pose a serious threat to human survival. However, this concern is not entirely accurate. Under reasonable control conditions, the integration of artificial intelligence and weapon systems may also increase the ethical dimension of future wars.

The concept of an "autonomous - firing artificial intelligence system" mainly includes two key components: "artificial moral agent" and "weapon". An "artificial moral agent" refers to an intelligent system with moral judgment and decision - making abilities. In the field of artificial intelligence, it can be achieved through learning algorithms and data models. However, it should be noted that current artificial intelligence systems, although having a certain degree of autonomy, still need human supervision and management to ensure that their behaviors comply with moral and legal standards. For example, in military operations, artificial intelligence systems can assist human commanders in making decisions through real - time analysis of the battlefield situation, but the final decision - making power still lies in the hands of humans.

The concept of "weapon" is relatively vague. From traditional crossbows to modern ballistic missiles, they can all be collectively referred to as weapons. Generally speaking, weapons are tools or devices used for attack and defense. In war, the use of weapons often causes great destruction and casualties, so certain moral and legal standards must be followed. International war laws clearly stipulate the scope and methods of weapon use and prohibit the use of some weapons with excessive lethality or the nature of killing innocent people.

The integration of artificial intelligence and weapons takes various forms. In the aspect of unmanned combat platforms, unmanned aerial vehicles, unmanned combat vehicles, and underwater robots have gradually been applied in the military field. These unmanned combat platforms can perform tasks in dangerous environments and reduce casualties. In reconnaissance missions, unmanned aerial vehicles can use their high mobility and concealment to obtain intelligence in enemy - occupied areas; unmanned combat vehicles can undertake tasks such as fire support and material transportation on the battlefield. In the intelligentization of weapon systems, by applying artificial intelligence technology to weapon systems such as missiles and artillery, the accuracy and strike effect of weapons can be improved.

Smart missiles can automatically adjust their flight paths according to the movement trajectories of targets and the surrounding environment to achieve more accurate strikes. Artificial intelligence can also be applied to battlefield situation awareness and command decision - making systems. Through the analysis and processing of massive data, it can provide commanders with real - time and accurate battlefield information to assist them in making scientific decisions.

However, the integration of artificial intelligence and weapons also has potential risks. Technical out - of - control is an important risk factor. Due to the complexity and uncertainty of artificial intelligence systems, there may be algorithm loopholes or hacker attacks, leading to the out - of - control of weapon systems. In cyber warfare, hackers may invade artificial intelligence - based weapon systems, tamper with their programs, and cause them to misidentify attack targets, resulting in serious consequences. The decisions of autonomous weapon systems may trigger ethical controversies. When weapon systems have autonomous decision - making abilities, how to ensure that their decisions comply with human moral and legal standards is an urgent problem to be solved. In some cases, autonomous weapon systems may misjudge the battlefield situation and cause harm to innocent civilians.

5.2. The Necessity of Developing Intelligent Weapon Platforms with Automatic Ethical Decision - Making Mechanisms

Developing intelligent weapon platforms with automatic ethical decision - making mechanisms can generally increase the ethical dimension of future wars rather than weaken it. This view is not groundless but is based on in - depth considerations from multiple aspects. It should be clear that "increasing the ethical dimension of future wars" does not mean eliminating wars. The outbreak of wars is often affected by a combination of complex economic, political, and cultural - psychological factors. Mere technological progress cannot fundamentally eliminate these underlying causes. From this perspective, the "unethical" or "immoral" nature of war itself belongs to the discussion scope of political science and political philosophy and is not the core of our discussion here.

Why can highly automated weapon platforms with "strong ethical functions" improve the ethics of future wars? This can be explained from the following key aspects. From the perspective of the war's humanity index, the "weak ethical attribute", "medium ethical attribute", and "strong ethical attribute" form a complete moral spectrum. In the process of weapon research and development, the gradual enhancement of ethical agency often indicates an improvement in the war's humanity index. Looking back at World War II, free - fall aerial bombs, due to the lack of enemy - friendly identification capabilities and precision - guidance functions, inevitably caused a large number of civilian casualties during use. During the Cold War, due to the limited accuracy of weapons, a large number of small - scale tactical nuclear weapons developed by the "Warsaw Pact" and "NATO" groups, such as nuclear torpedoes and nuclear artillery shells, would have caused huge damage to the environment if used. However, with the rapid development of modern technology, it has become a reality to conduct precise strikes on targets from a long distance with fewer ammunitions. This transformation of the war mode has greatly reduced the casualties of innocent people and environmental damage, fully demonstrating the positive impact of enhancing the ethical attributes of weapon platforms on the humanity of war.

From the development history of weapons, weapons that can fire automatically have long existed, and we should not be overly surprised by this. In traditional land defense battles, laying mines is a common tactical means to make up for the shortage of defensive forces. Mines, as the most primitive "automatic - firing weapons", will explode automatically once an enemy touches them. Compared with mines, automatic-firing platforms with ethical autonomy are not unacceptable from an ethical perspective. On the contrary, this proposed new type of weapon has the ability to select different levels of strike means according to different targets, thus being able to avoid more unnecessary harm. In some anti - terrorism operations, automatic - firing platforms can select appropriate weapons and attack methods according to the threat level of terrorists, which can not only effectively combat the enemy but also minimize the impact on civilians and the surrounding environment.

Some people may think that existing precision - guidance technologies can already meet the requirements of future wars for humanity, and there is no need to give weapon platforms the right to fire autonomously to avoid increasing unnecessary ethical risks. However, these views ignore the huge advantages brought by giving weapon platforms the right to fire autonomously, the most important of which is the great liberation of human resources. Human soldiers have physical and psychological limits, and the use of human soldiers also brings high human costs, which is an important bottleneck restricting traditional military force deployment. Weapon platforms with automatic - firing capabilities will greatly enhance the flexibility of military force deployment, and this advantage is also reflected in remotely - operated unmanned aerial vehicles.

Although current unmanned aerial vehicles rely on human resources for control to a certain extent, with the continuous development of technology, unmanned aerial vehicles with automatic - firing capabilities will be able to perform tasks more autonomously and reduce dependence on human resources. In special tasks such as preventing terrorist attacks, this flexibility has great tactical value. Under traditional combat conditions, human commanders often spend a lot of time reporting to their superiors. Especially in cross - time - zone operations, time - zone differences and the interference of human biological rhythms will further affect the reporting efficiency. Terrorists may take advantage of this to attack civilians at the time when humans are most tired. A large number of armed robots are not affected by the "human sleep rhythm" and can conduct round-the-clock patrols. This will effectively fill the defense gaps and enhance the ability to prevent terrorist attacks without increasing the human resource costs of the military and police.

Existing unmanned aerial vehicles (UAVs) are mainly used to attack ground targets and are relatively less applied in the field of air combat. If UAVs for air combat are to be developed, remote operation by human soldiers may not be practical. Although the powerful 5G communication technology can provide certain support, in a war environment, the 5G communication network of one's own side is very vulnerable to being damaged by the enemy. Compared with UAVs that attack ground targets, UAVs used in air combat require higher mobility and faster reaction speeds. Therefore, developing self - firing aircraft may become an inevitable trend in future air combat. The emergence of such weapons will not only not reduce the humanity of air combat but may instead increase it.

Since unmanned fighter jets can fully utilize their mobility without considering the endurance of the human body, this will prompt potential enemy countries to also develop similar technologies to avoid being at a disadvantage in military struggles. The result of this mutual emulation may lead to the full unmanned operation of future air combat, creating a new situation in air combat where there are wins and losses but no casualties. This "bloodless" situation will make the post - conflict handling of military conflicts easier and reduce the stimulation to the populist forces within each warring country.

5.3. Concerns about Weapon Platforms Rebelling Against Humans and Responses

Facing armed platforms with autonomous firing capabilities, people's concerns are not unfounded. One of the most concerning issues is whether they will rebel against human commanders. The act of "rebellion" is generally regarded as an overt defiance of superior orders by an individual with self-awareness. In human society, Lu Bu's rebellion against Dong Zhuo is a typical example. Driven by his own interests and desires, Lu Bu violated his loyalty to Dong Zhuo and committed an act of betrayal. However, we would not consider a hammer that accidentally hits the user's finger as "rebellious" because a hammer is a tool without self-awareness. It cannot understand the meaning of its actions, and its behavior is merely the result of physical movement.

In the field of artificial intelligence, to acknowledge the possibility of robot soldiers rebelling against humans, we must first admit that robots have the ability to generate "desires" that conflict with human commands. This possibility is closely related to the design and operation mechanisms of artificial intelligence. According to "Asimov's Three Laws," robots are required to follow human commands, protect their own "bodies," and execute a priori normative instructions. However, in complex and changeable practical application scenarios, conflicts may arise among these normative requirements. In a specific combat situation, A, the task of blowing up the enemy's command post (requirement "A") may cause harm to innocent people in the vicinity, which conflicts with the requirement of not harming the innocent (requirement "B"). At this time, the robot needs to judge which requirement is more urgent based on the specific situation, and this judgment may lead it to violate human military commands.

Take an autonomous-firing drone as an example. When it observes a large number of schoolchildren passing by the target, it may violate the human command to launch a precision-guided bomb and suspend its operation out of protection for innocent lives. In a sense, this "rebellious" behavior is not necessarily a bad thing. In this case, the military robot does not betray the human commander by following a completely different set of behavior norms. Instead, it has a disagreement with the human commander on the issue of "how to determine the priority sequence of conflicting norms." This indicates that if we can pre-imbue future robot soldiers with widely recognized military moral principles, such as not torturing prisoners and protecting civilians, then when individual human commanders issue commands that violate these principles for certain complex reasons, the "moral code" pre-set in the robots can come into play and serve as a defense against human violations.

Scenarios envisioned by techno-apocalyptic thinkers like Elon Musk, such as all robots uniting to rebel against human rule, are highly impactful but have a low probability of occurring in reality. Before seriously discussing this possibility, we need to deeply consider the motives for robots to jointly rebel against humans. Looking back at the history of human military conflicts, they often involve the scramble for various biological resources, such as land, rivers, and population, which are crucial for human survival and development. However, as "silicon-based entities," the operation of intelligent robots does not directly depend on these biological resources. They do not need land to grow food, nor do they need rivers to obtain water resources, and they do not require a population for production and reproduction. Therefore, from the perspective of resource requirements, it is hard to imagine that they would be interested in occupying fertile land or having access to irrigated rivers.

Of course, the operation of intelligent robots requires a large amount of traditional energy, such as electricity and oil, which may lead to competition between them and human society in the energy field. However, it remains uncertain whether the new energy consumed by the widespread application of intelligent devices can be offset by the corresponding reduction in human activities. With the continuous development of science and technology, there is also great uncertainty about whether the application of nuclear fusion technology in the future can once and for all solve the global energy problem. Given the uncertainty of these key factors, it is insufficiently based in reality to be overly eager to discuss the possibility of machines uniting to compete with humans for energy.

In essence, "autonomous firing ability" is not the same as "the right to start a war." "Autonomous firing ability" merely refers to the tactical ability of a robot to select targets and launch attacks based on its own perception and judgment of the battlefield environment after receiving an order from a human commander. In actual combat, the autonomous firing behavior of a robot is carried out under the overall combat intention of the human commander and needs to comply with the combat rules and instructions set by humans. The risk of robot soldiers breaking away from the control of human commanders is no greater than that of human soldiers disobeying their superiors, and in some cases, it may even be smaller.

This is because robots can be more precisely controlled and monitored through programming and technical means, while human soldiers may disobey orders or desert on the battlefield due to factors such as emotions and willpower. Those who attempt to exaggerate the ethical risks of robot soldiers might as well first consider the ethical risks of using human soldiers and then conduct a more comprehensive and fair assessment of the advantages and disadvantages of using robot soldiers and human soldiers. Through such an assessment, we can have a clearer understanding of the role and value of robot soldiers in future wars and provide a more solid theoretical basis for the research and development of AI - enabled weapon platforms with autonomous firing capabilities [1-7].

References

1. Llorca Albareda, J., Liedo, B., & Martínez-López, M. V. (2025). Trusting the (un) trustworthy? A new conceptual approach to the ethics of social care robots. *AI & SOCIETY*, 1-16.
2. Balahurovska, I. (2024). The Importance of Roboethics In Innovation Technology Management. *System Safety: Human - Technical Facility - Environment,6*(1):16-25.
3. Dobos, N. (2020). Ethics, security, and the war-machine: the true cost of the military. Oxford University Press.*Criminal Law and Philosophy,* 2023,17(3):759-764.
4. Brown-Gaston, R. D., & Arora, A. S. (2021). War and peace: Ethical challenges and risks in military robotics. *International Journal of Intelligent Information Technologies (IJIIT), 17*(3), 1-12.
5. Chavannes, E. (2019). Towards Responsible Autonomy: The Ethics of Robotic and Autonomous Systems in a Military Context. *Hague Centre for Strategic Studies*.
6. W P S. Military robotics and ethics: a world of killer apps. *Nature*, 399-401.
7. Buechner, J. (2018). Two new philosophical problems for robo-ethics. *Information, 9*(10), 256.