INTRODUCTION

Nobody ever defended anything successfully; there is only attack and attack and attack some more.

George S. Patton, Jr.

Research to develop powerful autonomous systems, artificial general intelligence (AGI), and in the future, possibly, superintelligence is predominant in various academic and policy-making spheres. These technologies could transform mankind, and the planet itself. Some believe AI to be the technological development that humanity needs to achieve cures for terminal diseases and end biological and earthly limitations, among other benefits—and perhaps more importantly, to overcome its historical anthropogenic contradictions, such as wars, injustice, and inequality. Some even argue this is the last invention we would ever need to produce, as once superintelligence is achieved, it will create incredible technological developments that our biological brains cannot even imagine. Of course, an existential risk could arise if superintelligence is achieved and decides that humans are no longer of use or interest. There is also the strong likelihood that AI will be used for war. Facing these probabilities, calls for AI aligned with good human values and to benefit mankind are plentiful. Nick Bostrom has recommended AI development under what he calls the common good principle: “[s]uperintelligence should be developed only for the benefit of all humanity and the service of widely shared ethical ideals.” Proposals for precisely which ideals or guidelines should regulate AI production are often linked to their potential use as weaponry. A conference on AI in 2017 developed a set of values called the Asilomar AI Principles to direct research toward creating AI for the common good, and Principle 18 explicitly states: “An arms race in lethal autonomous...
Artificial Intelligence Safety and Security

weapons should be avoided." A letter was also signed by thousands of AI researchers at the International Joint Conference on Artificial Intelligence in 2015 warning humanity of the potential dangers of militarized AI:

> Artificial Intelligence (AI) technology has reached a point where the deployment of such systems is—practically if not legally—feasible within years, not decades, and the stakes are high: autonomous weapons have been described as the third revolution in warfare, after gunpowder and nuclear arms.\(^9\)

Federico Pistono and Roman Yampolskiy have coined a term that applies to military use of AI: Malevolent AI (MAI). This MAI could exploit weakness in our nuclear weapons storage facilities or create a virus to act against us, but could also be used to achieve dominance.\(^10\) The use of MAI for dominance—henceforth the term hegemony will be used—will be a central factor of this chapter. Pistono and Yampolskiy are not alone: Elon Musk, Bill Gates, and Bill Joy are among many technologists who also think AI could turn against us.\(^11\) Toby Walsh, an AI scientist who often speaks at United Nations meetings on autonomous weapons, worries about terrorists or others with malevolent intentions using powerful AI to hurt us.\(^12\) Indeed, the very development of AGI and superintelligence could be the work of only a few unaffiliated (and thus unregulated) individuals sitting in a garage with their computers. However, in this chapter we are more concerned about important shifts in global security and the international system due to militarized AI. This is likely to result from the actions of great powers: historically, those that can destabilize the international system, causing deaths on a massive scale and the reproduction of an anarchic international structure. From the perspective of international relations (IR), the great danger of MAI in relation to global peace, security, and the preservation of humanity is its use not by terrorists but by great powers jostling for hegemony.

Therefore, our emphasis is on what impact the development of AI could have on the international system. Because various militaries are rushing to develop militarized AI to add power to their conventional weapons, we consider this to be an important focus of study.\(^13\) This chapter finds that calls for safe AI are valid and necessary, and in the far future they could provide us with the framework for friendly AI. However, in order to understand how, from the perspective of great powers, AI and militarized AI will play a critical role in the drive for hegemony, we need to place this topic within a wider field of abstraction in relation to global security.\(^14\) This chapter aims to enhance the discussion of the consequences of militarized AI by looking at it from the field of IR. The particular lens we will use is the theory of offensive realism (OR), which tells us that the international system is anarchic—this does not mean there is widespread chaos, but that the lack of a world government leads states to joust for power.\(^15\) In the anarchic structure of the international system, the only way to maintain the safety of the state is to have power, and power in IR means military power. Furthermore, OR avers that if a state has substantial power, it will aim to project it in order to achieve regional hegemony. This causes fear and insecurity in other great powers, creating a security dilemma in which states perceive each other's defensive measures as potential threats.\(^16\) Increased militarization by great powers affects the entire globe, and currently the great powers of our world—the United States, China, and Russia—are those most active in pursuing militarized AI. John Mearsheimer, the main proponent of OR, maintains that great powers can only aspire to regional hegemony (not global), because powers that are separated by great bodies of water “usually do not have much offensive capability against each other, regardless of the relative size of their armies.”\(^17\) This chapter argues that powerful AI may change that: if it can provide great powers with technological superiority over other great powers, it could be used offensively to obtain global hegemony. This will be the only way to secure survival in an era of rapidly evolving security dilemmas that engulf all possible geographical and technological spheres.

Mearsheimer’s theory of OR will serve as the framework for this chapter, providing a broader level of abstraction in which to place the implications of militarized AI in relation to great powers and drives toward hegemony. First, the chapter presents the theory of OR and what it tells us about
Offensive Realism and the Insecure Structure of the International System

the international system, hegemony, and the maximization of power. Second, it presents a panorama of the current security dilemma, militarized AI Research and Development (R&D), and plans for the future. Third, this chapter analyzes various cases based on AI development scenarios to present a series of predictions of how militarized AI could be used by great powers, and whether militarized AI could provide great powers with the possibility to break the stopping power of water, allowing them to drive toward global hegemony. This would herald an era of constant and growing security dilemmas contrary to the global greater common good, as well as possible existential risk.

**OR AND THE INSECURITY OF THE INTERNATIONAL SYSTEM**

General “Buck” Turgidson: Sir … he’ll see everything. He’ll see the big board!

Dr. Strangelove: The whole point of the doomsday machine is lost if you keep it a secret! Why didn’t you tell the world, eh?

“Dr. Strangelove” Stanley Kubrick

Though there are many fields working on AI and its impact on humanity, IR is still not fully immersed in this new area of study. This is a problem if we want to understand how other technological developments have affected societies socially, economically, and politically, and how the nature of the international system affects global security.18 Forecasts estimate a 50% chance that we will have high-level machine intelligence by 2040–2050, and a 90% chance it will occur by 2075.19 Although these estimates are debatable and, in part, aimed at promoting informed discussions on the subject, it is telling that a large number of AI experts predict a promising possibility for human-level machine intelligence; thus these forecasts should not be simply ignored. After this, the speed at which AGI will transcend to superintelligence might be intensely accelerated.20 The incredibly rapid technological advancements anticipated from AGI and superintelligence, including militarized AI, force us to link them to an old IR question: what is the nature of the international system? This question is not purely academic, and it could help us predict the future of the international system as great powers aim to gain militarized AI supremacy. At present, drives to develop more powerful military hardware continue, and military expenditure increased worldwide in 2016 from the previous year.21 In a world where greater efforts are made to improve global governance through international institutions, why do states continue to arm themselves in this manner? OR can help us to answer this question—and to anticipate and analyze the future struggles for hegemony as AI becomes more powerful.

OR departs from five core assumptions about the international system. The first is that the lack of a world government (such as a Singleton,22 as proposed by Bostrom) makes the international system anarchic. The second is that great powers have offensive military capabilities, and states become more dangerous as their power increases; a state’s power supremacy is linked to some specific weaponry. The third is that states cannot be certain about other states’ intentions:

[N]o state can be sure that another state will not use its offensive capability to attack the first state. This is not to say that states necessarily have hostile intentions. Indeed, all of the states in the system may be reliably benign […] Uncertainty about intentions is unavoidable, which means that states can never be sure that other states do not have offensive intentions to go along with their offensive capabilities.23

The fourth assumption is that great powers have survival as their primary goal. The fifth is that states are rational actors, and they understand that their behavior affects the survival of other states and vice versa.24 The imbrication of these five assumptions leads states to act offensively toward each other, generating three patterns of behavior: “fear, self-help, and power maximization.”25 This is not a positive view of the world, and it does not mean to be. But the Realists argue that it provides us with a more accurate view of the world. In *The Tragedy of Great Power Politics*, Mearsheimer analyzes the shifts of power from the nineteenth century to the twenty-first century and proves that
great powers, including Imperial Japan, Nazi Germany, the United States, and the Soviet Union, have all tried to gain regional hegemony—a struggle in which the United States has successfully dealt with all rivals. This chapter aims to use this pessimistic view of the international system as a framework in which to analyze how the development of AI could result in the intensification of drives for hegemony by great powers. Using OR as our theoretical base allows us to account for a worst-case scenario by unpacking a warning of what could come to pass.

It is key to emphasize that all types of weapons create what is known as a security dilemma. This results from the increase of security (the production or purchase of weapons and deployment of power) by a state, perhaps increasing its naval fleet or number of nuclear warheads or the association with other states to obtain strategic advantages, which leads others to do the same, and this leads to a cycle of accumulation of power that can end in a confrontation that none of the actors really desire. Thus, the accumulation of any military technology, whether it be aircraft carriers or AI, leads to a security dilemma. Mearsheimer argues that conventional forces increase fear and insecurity in great powers more than nuclear weapons, as nuclear deterrence makes a confrontation less likely, although not unthinkable. Mearsheimer explains:

[R]ival states that possess nuclear forces that can survive a nuclear attack and retaliate against it are likely to fear each other less than if these same states had no nuclear weapons. […] The logic here is simple: because nuclear weapons can inflict devastating destruction on a rival state in a short period of time, nuclear-armed rivals are going to be reluctant to fight with each other, which means that each side will have less reason to fear the other than would otherwise be the case.

This is concomitant with the stability–instability paradox, where states that possess nuclear weapons are less likely to face each other directly, but do it indirectly through smaller conflicts. The numerous proxy conflicts that arose during the Cold War illustrate this, and the horrors caused by proxy wars in countries such as Vietnam, Cambodia, El Salvador, Nicaragua, Chile, Angola, and Colombia, to mention a few, will live in the collective memory of mankind as some of the most horrendous crimes we have experienced. Whether militarized AI will herald a similar era is something we will address here.

OR’s third assumption is that a state cannot be certain of its rivals’ intentions. David Sorenson and Michael Wolfson argue that competitors will increase their military R&D in order to minimize uncertainty, and this accumulation of power can lead to escalation. However, states can know with some degree of accuracy their rivals’ capabilities. Knowledge of capability drives development of arms and weapons platforms in order to stay competitive, rendering the security dilemma ever more dynamic and unstable. This was illustrated by Stuart Armstrong, Nick Bostrom, and Carl Shulman, who proposed a model of an AI arms race between several AI teams. One of the key conclusions was that it would be better for AI teams not to know their capabilities, as “increasing the information available to all the teams (about their own capability or progress towards AI, or that of the other teams) increases the risk [of an arms race].” This corroborates the OR observation that “[w]hen a state surveys its environment to determine which states pose a threat to its survival, it focuses mainly on the offensive capabilities of potential rivals, not their intentions.” One of the variables of the model proposed by Armstrong, Bostrom, and Schulman is enmity, which they set equal for all. The authors write, “[e]nmity is something that we can work on by, for instance, building trust between nations and groups, sharing technologies or discoveries, merging into joint projects or agreeing to common aims.” For OR, however, the issue of enmity counts little in terms of the drives for hegemony. Even a benign power’s intentions will be doubted by others, as it is impossible to know with 100% certainty the intentions of the first state, especially as it is assumed that all states want to survive in the anarchic international system. Thus, uncertainty about the competitors’ intentions and also knowing their capabilities increases the likelihood of military buildup and conflict.

A key contribution of OR to help us predict possible consequences of powerful militarized AI is the concept of regional hegemony. OR conjectures that, in a self-help world, great powers
will aim to increase their power to such levels that they can be the regional hegemon. The United States is the regional hegemon in the Western Hemisphere, as it has no potential rivals there. If a great power achieves regional hegemony, it can roam far from its neighborhood to prevent other powers achieving hegemony in their own regions through the maximization of power. However, Mearsheimer maintains that there has never been a global hegemon, and it is unlikely to happen any time soon because of the great obstacle posed by the stopping power of large bodies of water. Conventional deployment of hardware and human forces is constrained by biological and current physical limitations. Militarized AI that can be permanently deployed and can knock down the rival’s cyber systems, AGI, and beyond could break the stopping power of water. These technologies will vastly increase the theater of operations, from cyber to outer space; thus the sphere from which an attack could take place is enlarged significantly due to militarized AI having the ability to encroach in areas that were off-limits before—when regional hegemony was disputed via traditional hardware. Here we borrow from Richard Harknett and Michael Fischerkeller to add an extra layer of analysis to OR. They state that such new technologies as cyber and, in the future, machine learning and AI will change the nature of deterrence. These technologies will allow actors to relentlessly search for the initiative, resulting in what they have called a “condition of constant contact” and “cyber persistence” between rivals. This implies highly frequent encounters between forces due to the advancement of AI and cyber systems, meaning that cyber deterrence is extremely unlikely. The advance toward more powerful and ever present militarized AI would, therefore, herald an era of constant and growing security dilemmas, contrary to the global greater common good. This could be worse if militarized AI provides significant qualitative advantages—and most literature on AI predicts this will be the case. Michael Intriligator and Dagobert Brito argue that qualitative arms races are more unstable than quantitative, because superior arms create the opportunity to strike first. This allows us to add another focus of scrutiny to our arguments, underpinned by the three resulting behaviors from the five core assumptions of OR (self-help, fear, and maximization of power), and enhanced by the nature of a qualitative arms race and the condition of constant contact. Our view is that the combination of these realities will create an AI security dilemma in which great powers are faced with shrinking regional hegemonic possibilities, driving them more aggressively toward global hegemony—breaking the stopping power of water to secure their existence. This could result in an era of vast militarized efforts and a dystopian scenario for mankind, contrary to a dreamed greater global common good where friendly AI provides mankind with democratic, technological, and intellectual abundance.

CONVENTIONAL AND AI SECURITY DILEMMAS

But because it was the very world it was, the very world they had allowed it to become, for months his activities did not come to the alarmed attention of The Ones Who Kept the Machine Functioning Smoothly.

Harlan Ellison

THE CURRENT SECURITY DILEMMA

The current U.S. defense budget is US$619 billion dollars, and the requested budget for 2018 is US$639 (including the budget for overseas operations). This is substantially more than its main rivals, Russia and China, which spend US$69.2 and US$215 billion per year, respectively. The year 2016 saw weaponry record sales higher than in the last years of the Cold War. The United States is now more fearful about China than it was a few decades back, and this fear will continue to increase as China develops weaponry that could pose a challenge to U.S. supremacy. For example, both China and the United States are investing heavily in new aircraft carriers. The United States has 10 aircraft carrier battle groups deployed throughout the world; yet, other powers such as China are trying to gain some ground in this strategic area, with two aircraft carriers of lesser tonnage being
deployed. These present no match for U.S. carriers, which confer a qualitative advantage by having superior tonnage, accommodating more planes, and having more advanced technology. China is planning to strengthen its naval capabilities to achieve what has been called the Blue Water Navy—a program to “show the flag” far from the mainland, and especially in the South China Sea, where China has various border conflicts with its neighbors, including Japan, the Philippines, and Thailand among others (all U.S. allies). To maintain superiority in this area, the United States is aiming to replace its Nimitz-type carriers with 10 newer Gerald Ford-type carriers, which can pack more power than rival carriers. The Gerald Ford-type carriers also incorporate automation technology, thus requiring less staff to run them, and have complex automated self-defense systems. Former State Department Secretary Hillary Clinton pinpointed the U.S. preoccupation with the rise of China in Asia by outlining what she called the Pacific Century, in which it is evident the United States has no plans to lose its power position in Asia. China’s very active military program is a response to this. This cycle epitomizes a security dilemma. In Destined for War: Can America and China Escape Thucydides’s Trap?, Graham Allison posits that a war between China and the United States is likely. Using the Thucydides Trap as conceptual base, he describes how the rise of China will challenge the incumbent power, the United States, which will result in conflict. Mearsheimer similarly avers that the United States does not tolerate competitors, and that a future clash with China is likely. These are worrying predictions, as these two powers are both aiming to obtain a significant power advantage by developing militarized AI.

Although this section is brief, it illustrates that a conventional arms race continues. Currently, any conventional weapon can simply blast the most advanced AI system into smithereens. Thus, the level of destructive power of conventional weapons is not to be ignored. On the other hand, as the United States, China, and others focus more on developing militarized AI, we must wonder what the international system will look like once this goal is fully realized.

AUTONOMOUS SYSTEMS: FUELING THE SECURITY DILEMMA

Weaponry accumulation and security dilemmas do not cease; they just become transformed by technological advancements. Beyond the accumulation of conventional military capabilities, great powers will aim to obtain a significant advantage in AI development.

The United States has no intention to lose the race on militarized AI. The long-term strategy is to maintain military supremacy in what former Secretary of Defense Ashton Carter referred to as the Third Offset: an evolution from purely conventional power to a more refined focus on the next weaponry that can offset rivals’ capabilities. This means looking into artificial intelligence and autonomous weapons that will provide power supremacy to the United States. The Third Offset aims as stated by former Deputy Secretary of Defense Robert O. Work are to “achieve comprehensive stability, reduce any incentive for preemption, and if we do come to blows, end it quickly before we trip over a nuclear threshold.” A plan for the next 30 years highlights this emphasis on U.S. autonomous weapons development, and the necessity of being at the forefront of autonomous systems: “[u]nmanned systems and autonomous software offer significant potential advantages for meeting the challenges of a newly forming adversarial environment.” General Paul J. Selva of the Air Force, and vice chairman of the Joint Chiefs of Staff, recently commented that the United States is within 10 years of creating a fully autonomous system that could decide when and whom to kill, although they had no interest in building one. He added that other countries are not far behind, and that someone would unleash it, eventually. If a devastating AI weapon is created and used, it is likely that others will advance research to obtain it in order to maintain or achieve hegemony. The race for nuclear weapons exemplified this, even before one was used in war.

Who is ahead in this arms race is difficult to state. Yet, the United States has had to consider the possibility of not having overall power supremacy in autonomous systems. At the Space Symposium in 2016, Robert O. Work stated in relation to China and Russia:
[W]e would once again be forced to contend with the re-emergence of great power competition. The resurgence of Russia and the rise of China will require that we exercise strategic muscles that we’ve allowed to atrophy since the end of the Cold War. [...] And both countries are not just getting good in the usual domains of air, land, and sea, but also especially in cyber, electronic warfare, and space. As a result, our margin of technological superiority is slowly eroding, and addressing this issue is one of our most important strategic tasks, because too great an erosion of our technological superiority would ultimately undermine our conventional deterrence, raise a competitor’s incentives for preemption, contribute to crisis instability, and greatly raise the potential cost of any future U.S. military operation.51

Defense Secretary Jim Mattis commented at a Senate hearing that: “[f]or decades, the United States enjoyed uncontested or dominant superiority in every operating domain or realm. We could generally deploy our forces when we wanted, assemble them where we wanted and operate how we wanted. Today, every operating domain—outer space, air, sea, undersea, land and cyberspace—is contested.”52 Were Russia, China, or other powers to gain in AI superiority that could not be countered with the advantage the United States has in conventional hardware, it could signify the rise of a new regional hegemon. Although Russia is rushing to develop powerful autonomous systems,53 it is China that is probably going to challenge the United States in this arms race.54 Under President Xi Jinping’s vision to form a civil–military partnership on AI R&D, China hopes to develop AI power capabilities to surpass those of the United States, including some that aim to counter powerful U.S. weapons platforms.55 A statement from Lieutenant General Liu Guozhi, director of the Chinese Central Military Commission’s Science and Technology Commission, highlights not only the core premise of OR but also the condition of constant contact, and thus the necessity to achieve powerful militarized AI before their rivals: “Whoever doesn’t disrupt will be disrupted!”56

Within the AI literature, the militarization of autonomous systems might be construed as MAI; indeed, any technology used for killing might fit this description. However, OR tells us that states are rational, and thus maximizations of power are not simply isolated, irrational, and malevolent occurrences but rational initiatives that can be understood from a broader level of analysis. These offensive drives make sense within an international system where fear and insecurity, and the maximization of power, are not the anomaly but normalcy. AI scientists state that commercial R&D is well ahead of the military’s, and therefore AI R&D open source foundations could secure good AI. The basic line here is that commercializing AI and also banning autonomous technology for military application will prevent a dystopian scenario. A Chatham House research paper argues that banning militarized AI may not be efficient, as superior technology will be available in the commercial sector.57 Indeed, if powerful AI technology is available in the commercial sphere, it is unlikely that states will be impeded from gaining hold of it. It might take significant international treaties to prevent militaries worldwide using it, which still might not be enough if history tells us anything—international norms on climate change are still affected by considerable difficulties owing to economic and ideological pressures, as are treaties on nuclear weapons.58 Until smoothly running global governance is achieved, which perchance might never take place, we can expect struggles for hegemony to continue into the far future.

EVOLVING AUTONOMY: BREAKING THE STOPPING POWER OF WATER

There are no secrets about the world of nature. There are secrets about the thoughts and intentions of men.

J. Robert Oppenheimer

In the anarchic world of international politics, it is better to be Godzilla than Bambi.

John Mearsheimer59

We have established that a nascent AI security dilemma is in motion. Great powers are rushing to develop powerful AI systems in order to gain commercial and military superiority. Currently, AI military technologies still operate on a rules-based level. This includes everything from the US$1
trillion dollar F-35 stealth manned fighter project described by former U.S. Deputy Secretary of Defense Robert O. Work not as a plane but as a “[…] flying sensor computer that sucks in an enormous amount of data, correlates it, analyzes it and displays it to the pilot on his helmet” to the cheaper unmanned aerial vehicles (UAVs), including the autonomous U.S. navy fighter/bomber X-47. Though impressive in their own right, these technologies still cannot reason, which would require knowledge-based behavior. They can gather and relay information faster than humans, but they are still more automated than autonomous.

Timeline predictions on AGI and superintelligence development suggest that in the near future these weapons systems will be able to perform complex and fast operations, matching our biological brains or better, providing a qualitative power advantage that could represent a more immediate and far-reaching threat to great powers. U.S. historian Alfred W. McCoy presents such a scenario from his analysis of previous and current U.S. military developments:

It’s 2025 and an American “triple canopy” of advanced surveillance and armed drones fills the heavens from the lower- to the exo-atmosphere. A wonder of the modern age, it can deliver its weaponry anywhere on the planet with staggering speed, knock out an enemy’s satellite communications system, or follow individuals biometrically for great distances. Along with the country’s advanced cyberwar capacity, it’s also the most sophisticated militarized information system ever created and an insurance policy for U.S. global dominion deep into the twenty-first century.

This leads us to envision a situation where autonomous systems, some in development at present, can hunt rival capabilities, avoid counterattacks, modify missions, and prioritize targets while in deployment, which could pose severe threats to battle groups, aerial vehicles, cyber systems, and defense environments highly supported by AI. The objective is to have these AI systems deployed for years and reach most places on the Earth, cyber, and the exo-atmosphere at hypersonic speeds without the need for human presence or constant input. Evidence suggests the United States is expecting that in the next major war, space capabilities, such as swarm of space nanodrones, and cyber capabilities enhanced by machine learning, will be key to maintaining supremacy. Great powers’ leaders will have a view of the theatre of operations, just like Dr. Strangelove had in the War Room in Stanley Kubrick’s dark comedy masterpiece. The U.S. Defense Advanced Research Projects Agency (DARPA) has already sought to develop a highly intelligent system, Deep Green, which would support strategic decisions in the theatre of operations through predictive capabilities, such as assessment of options and impact of decisions in real time. These types of real-time operational intelligent environments, improved by ever more sophisticated autonomous systems, will make the theatre of operations so large and complex that regions where states had hegemony will be erased, and a first strike could come from any sphere in which capability advantage could be had. This will increase the level of fear in great powers.

Recently, a U.S. warship passed within 12 nautical miles of a Chinese-built reef in Spratly Islands, South China Sea. Encounters of this type can raise already high tensions between China, its neighbors, and the United States. Although moderately infrequent now, these incidents could become a very common occurrence as autonomous planes, submarines, boats, and swarms of drones are deployed, in addition to cyber elements in the fashion described by McCoy. Here we revisit the concepts developed by Harknett and Fischerkeller, who argue that as technology advances, especially AI and machine learning, it is expected that rival powers will be continuously trying to gain the initiative in this nontraditional theatre of operations, and thus persistent engagement or cyber persistence is the only way to advance a power’s security. This cyber persistence is needed in order not to lose the initiative due to the evolving nature of cyber and machine learning. This creates a “condition of constant contact.” Unable to successfully secure their regional hegemony, great powers will opt to improve the qualitative nature of their power to deploy it offensively. This would increase fear among great powers, generating a very dynamic and precarious security dilemma, as attacking will be the only strategy that can guarantee supremacy.
Soon missiles will be more than exploding bombs; they will assess and develop strategic striking capabilities versus the defensive hardware of other powers, redeploy and build, and analyze theatres of operations at enormous speeds. Ultimately, the purpose of weaponry is to be effective in destroying the rival's hardware, human operatives, and various bits of infrastructure. The increasing sophistication and speed at which progress is made on AI R&D will be vital to secure supremacy in the international system. Following OR’s third assumption, great powers possessing powerful AI will still not be able to know the intentions of their rivals. This uncertainty will lead to increasing militarization, in this case a never-before-experienced qualitative arms race that could plunge the world into a dark era of widespread conflicts and proxy wars. If powers still keep their nuclear deterrence, this fragile balance created by ever-evolving militarized AI could take us back to Cuba 1962.

As AGI advances toward superintelligence, what scenarios can we envisage in relation to the international system? We assume we are still in a self-help world, in addition to the condition of constant contact underpinning powerful and disruptive drives for hegemony. Supposing militarized AGI has been achieved and has similar intelligence to humans, this momentous development can disrupt the international system if it can provide a power advantage against rivals—a weapons system that can define an era, as nuclear weapons did after World War II. We assume here that, despite the computing power of AGIs, humans are still in control—which is closer to the scenario the Pentagon envisages for the future of autonomous systems, where humans and AI work together in what they have conceptualized as the “centaur model.” The possession of AGI capabilities, just like power capabilities now, does not mean that there will be an imminent war, but an AGI security dilemma could escalate very rapidly, as AGI military technology allows powers to deploy their capabilities for longer, higher, and faster, and matching human intelligence.

In relation to AGI and superintelligence, we could propose some situations to assess if they do allow a great power to become a global hegemon. For this we turn to some of Bostrom's considerations on the speed of the transition from AGI to superintelligence, and possible consequences. Adding superintelligence to the discussion adds complexity and the possibility of existential risk for great powers. It is also important to remember that this is taking place at speeds beyond those of our biological brains, and that any humans involved in this might have artificially enhanced intelligence or perhaps exist as brain emulations, which could significantly increase their intelligence. Bostrom tells us that there are three possibilities for the transition between AGI and superintelligence: slow, moderate, and fast takeoff. A slow takeoff would occur over a long period of time, perhaps decades or centuries, allowing for political processes to adapt and create a strategy for how to deal with superintelligence. A moderate takeoff could take place over months or years. This scenario would not allow humans enough time to develop proper institutions, accords, and plans for contingency and could result in geopolitical, economic, and social conflict. Lastly, a fast takeoff could happen in minutes, hours, or days, not allowing for global coordination to take place or other types of responses to be put in place.

Let us first consider a scenario where there is AGI parity between powers, and they are using these capabilities to solve global governance issues but at the same time to increase their power capabilities. It is likely that no international treaty could effectively prevent drives for hegemony, assuming the international system remains anarchic, knowing that superintelligence may soon be achieved. Moreover, let us remember that power competition in this era is persistently contested. The ability to maintain or gain regional hegemony, if not gone entirely, could be a constant and fast changing tit for tat. There would be no way to tell when the first power was transitioning from having AGI to having superintelligence. If a state achieved this after a fast takeoff, it could be an existential threat to other powers; thus there is the likelihood that intensified arms races and conflicts would occur beforehand. If there is parity among powers, it is unlikely that the stopping power of water can be broken in order to achieve global hegemony. A more probable result would be permanent confrontations between great powers, akin to the ones that took place during the Cold War, but perhaps more destructive, as AGI capabilities would destroy the enemies’ cyber infrastructure and
Artificial Intelligence Safety and Security

meddle with economic, political, and health systems among others. Despite these smaller conflicts, great powers would still aim toward global hegemony as long as there was the latent possibility of superintelligence being achieved. It is more likely that a moderate takeoff, as Bostrom points out, could generate these turbulent times. However, if the international system remains anarchic, it is very possible that even a slow takeoff might be faster than the transformation of mankind toward greater levels of social technological development.

The reaction to a world where great powers are on permanent offensive mode will drive the sophistication of AI even further. An arms race that could negate the stopping power of water would instill more insecurity and fear in great powers—possibly leading to preemptive war. History has shown us that even international law can fall victim to a great power set on using this doctrine to secure its existence or further its interests. The speeds at which power development could take place could make the doctrine of preemptive war more common. This would signify a visible increase in the historical contradictions that we face now, as well as an AGI security dilemma that could run out of control.

A great power wielding powerful AGI capabilities that can transcend the biological and geographical constraints we have now will not stop at regional hegemony. Obtaining global hegemony will make this hypothetical great power very safe, as there will be no power that could threaten its existence. Since the end of the Cold War the United States has had power supremacy, albeit not as a global hegemon, and this has made it the safest of any great power in history. This has not stopped the United States adding more power to its arsenal, and it is unlikely that it will stop, as various statements have shown here that it perceives other states will try to challenge its regional hegemony. If the United States had enough AGI or superintelligence power capability to overcome the stopping power of water, it would try to gain global hegemony and ensure its security. If superintelligence is not constrained by biological limitations, for example, in the form of millions of nanodrones or brain emulations, then the stopping power of water will do nothing to prevent a great power driving toward global hegemony. Again, a great power would need to possess AI power capabilities sufficiently more advanced than others to achieve this. A permanently contested regional hegemony could make this scenario unlikely. If it occurred, great powers would intensify their militarized AI development, or another type of qualitative weaponry, to reduce uncertainty but increase fear in other powers—further intensifying the security dilemma.

Another scenario would entail that a power with AGI had (as they developed/evolved) that extra computational or intelligence power to solve mankind's global governance problems. This assumes that mankind was always close to solving them but was constrained by its biological limits. These new international norms could stop the creation of a superintelligence that could design a world government, a Singleton perhaps, to address humanity's historical anthropogenic contradictions; this might threaten great powers’ ability to gain global hegemony. Therefore, global governance as we conceive it now might not be enough to change the self-help system, and thus great powers might opt to remain in an anarchic system in order not to face an existential risk.

Nothing prevents great powers from using their AGI to work on improving global governance, perhaps achieving new ways to provide global solutions that human intelligence has overlooked. A caveat here is that we think global governance problems that we experience now can be resolved by our current intelligence—it is not the lack of biological computing power that is stopping us solving world problems, but how we have used this intelligence to form technological developments such as democracy, our international political economy, and ideologies that are adopted by competing forces.

It is argued that openness of AI development would allow more than the two most likely superpowers of the twenty-first century (the United States and China) to develop powerful AI. Bostrom argues that a multipolar world might make it harder for AI systems to cause “extreme damage […] If AIs are copies of the same template, […] they might all contain the same control flaw.” This assumes to a certain point that there could be a flaw shared by all AI systems, and this means all actors will have the same problems and advantages. Setting aside any flaws, historical appreciation would tell us that a multipolar world where many powers possess powerful AI is not as
safe as a unipolar or bipolar world. Numerous great powers with offensive AI capabilities, locked in a security dilemma, underpinned by the constant contact condition, unfolding at enormous speeds and in many theatres of operations will increase insecurity and fear in great powers and intensify offensive behavior. Efforts to avoid an AGI/superintelligence could work if its development could be controlled or regulated in such way that none of the parties could achieve superiority. This would mean a total openness on AI development. The openness of AI development has been discussed in a paper by Bostrom, yet it seems that it all comes down to global coordination problems:

A world in which global coordination problems remain unsolved even as the power of technology increases towards its physical limits is a world that is hostage to the possibility that – at some level of technological development – nature too strongly favours destruction over creation.\(^2\)

Throughout our history the international system reflects that achieving global governance is not a final static outcome but a series of complex ongoing processes, as states, including great powers, compete within the existing international political economy for economic and military supremacy. To change the international system and avoid an AI arms race toward hegemony, a reinvention of our international political economy is needed; however, this challenge lies outside the scope of this chapter.

In a self-help system, how AGI is developed and deployed could increase fear in great powers. This fear will lead great powers to maximize their power, focusing on qualitative advantages, which would deepen the fear factor, leading to a condition of constant contact; thus spins the cycle of self-help, increasing qualitative capabilities, fear, deployment of power, more fear, and so on. The only strategy for great powers to secure their survival will be to push for global hegemony. Even a reluctant great power in relation to global hegemony will have to aim for it in order to stop AI security dilemmas and ensure its own existence. This vision of the future presupposes that, as security dilemmas become more unpredictable, more dynamic, more aggressive, and as rapidly devastating evolving conflicts fill most spheres of life, mankind and our AI counterparts will be set back in their march toward a greater global common good.

**CONCLUSION**

If the iron dice must roll, may God help us.

Theobald von Bethmann-Hollweg  
*German Chancellor, August 1, 1914*

This chapter does not claim to be an exercise in soothsaying, nor can any other work on the future of mankind and AI. However, this does not stop us trying to seek speculative intellectual constructions on how to approach this very powerful and promising technology. Researching AI and its possibilities fills us with significant hope for what we could achieve as a global community. It is an area of study that brings with it a certain flare of positive charge. However, when we look at it from the discipline of IR, the possibilities of AI development start to look increasingly dark. The aim of this chapter was to contribute to the discussion on AI and global security from the perspective of great power politics. We wanted to take a wider view of how such a powerful technology would be used by great powers, regardless of the countless calls for good AI that fill the extant literature. We do not reject these calls, but instead propose to add an extra explanatory layer to contribute to this field. If we use the lens of OR, we can see not only what the future could bring but also why exactly it is important to emphasize this approach, as well as the contribution of IR to AI development, security, and society.

From the body of work on friendly AI we can draw some hope that at least there are efforts to resolve safety issues and build good principles for AI development. Nonetheless, calls for good AI have a substantial caveat, and this is the **political problem**. With the theory of OR we addressed part of the **political problem**, viewing the past, present, and future as less than idealistic. Of course, we
could not approach every single aspect of great power politics and AI in this chapter, but what we have presented provides us with certain clues about what the future could look like if militarized and powerful AI achieve their informed speculative potential.

We did not aim to explain how the power of AI could generate wars due to the irrationality and foolishness of leaders, the evil nature of man, domestic politics, nor the weak power of ideas. Instead, we proved that within the anarchic structure of the international system, the substantial qualitative power that AI could provide will result in an epoch where regional hegemony is permanently contested by constant offensive behavior, as AI allows great powers to reach all spheres of operations. This will take the current self-help world to an extreme in which uncertainty drives maximization of ever-improving AI capabilities, and the condition of constant contact is an added normalcy to the anarchic system. Due to this, gaining regional hegemony might become impossible, as great powers are bogged down with endless conflicts occurring at vast speeds. In response, the desire for human-level machine intelligence or superintelligence will be accelerated toward gaining a decisive power advantage to achieve global hegemony, and breaking the stopping power of water in order to guarantee the survival of the winner of the race toward the most advance AI achievable.

In all projected scenarios, when great powers become engaged in a condition of constant contact, or if a great power does achieve AI that allows it to achieve global hegemony, mankind and its AI cohabitants will be entombed by recurrent setbacks and prevented from achieving a greater global common good. The horrendous outcomes of the era of proxy wars during the Cold War would pale in comparison to the scenarios presented above. And every time it will be worse, as great powers continue to maximize their power, driving this macabre cycle toward existential risk.

If superintelligence is finally achieved, this would represent an epochal change, and we would hope this would mean that mankind’s historical anthropogenic contradictions would have been overcome, including the anarchic structure of the international system. This would also mean that the drive for hegemony would have disappeared. This might occur due to superintelligence itself, a conjecture based on the idea that superintelligence is capable of technological feats beyond our imaginations. However, if powerful militarized AI is developed within the current international security structure, where arms races, security dilemmas, and military aggressive hegemonic behavior are still very much the norm of international politics, the nightmare scenarios presented here are more likely to come to pass. Moreover, the achievement of superintelligence could become the last act of a tragic play. Its achievement would signify that most of mankind, AI beings, other living beings, and our pale blue dot had been victims of unimaginable conflicts. Under these circumstances, achieving superintelligence would leave a victorious great power to contemplate the lost opportunity for a new dawn in mankind’s history. This superintelligence could be everything we have dreamed it to be, but its victory would be hollow if the collective richness of our earth had been destroyed.

If we want to avoid this dystopian future, we need to include the somber essence of IR in our discussions of AI and the future of mankind. From this, we can contribute to other fields in determining good avenues toward the creation of AI for the greater global common good. A first step would be to start thinking about the creation of AI not solely as a way to enhance our existing aspirations, but as a way to resolve humanity’s historical anthropogenic contradictions, such as the need for the production and accumulation of military power, or the continuum of global inequalities, injustice, and a flawed international political economy that even in the face of a global environmental catastrophe increases its destructive cycles. If we leave it to AI futures to solve this, mankind will have failed as the steward of our planet and the only known being aiming to create superintelligence in the universe.

ENDNOTES

1. This chapter will use the term AI to include narrow AI, machine learning, artificial general intelligence (human level machine intelligence), and superintelligence, unless otherwise specified. Although this might seem arbitrary, it allows for clarity. All these powerful technologies can provide technological
superiority in relation to military power. Nonetheless, there will be some sections within the chapter where the distinction will be made between them. A variety of opinions on when, or if, we will have AGI and superintelligence can be found in academic literature and in the media, as everyone from AI scientists to technologists has weighed-in on this topic. Probabilities assigned to its creation vary greatly, but the idea of an intelligence superior to ours is clearly in the minds of scientists, being mused over as a realistic possibility. It is important to note that probabilities do not mean that humanity will achieve its aims, or will achieve them within the prescribed timeframe, but they serve as an indication of possibility, and this is enough for the main argument of this chapter, which aims to present possibilities based on current research.


3. If speciation were to take place, there is the possibility that post-humans would shape the planet to their needs as we have done, ourselves. How this would look, it is hard to know, but if the literature is a sign of possibilities then the ecological impact could be very different, as superintelligent beings might not have the same type of biological requirements that we do. If one of the results of the pursuit of AI is brain emulations, it is very possible these will decide to leave the planet altogether, perhaps allowing it to recover from the Anthropocene. Contrary to this, if AI and superintelligence maximize exponentially the exploitation and accumulation of resources and the drive for dominance, the odds for the planet to continue to hold biological life are low. Although the concept of brain emulations might be controversial among certain scientific spheres, for the purpose of this chapter the failure to achieve them does not undermine its main argument. Brain emulations represent one type of AI among many options that are being pursued. Elon Musk’s Neuralink at the very least highlights an interest in the idea of machine and human brain connection, and further on, brain emulations.

4. We contend that humanity’s *historical anthropogenic contradictions* such as wars, injustice, global inequality, and constraining challenges to freedom of intellect are the biggest challenges in relation to the future of AI development. The backbone of the author’s ongoing research argues that if AI is conceived as an extension of the current international political economy and not as way to overcome it, mankind’s *historical anthropogenic contradictions* will be carried into the future. This would result in a dystopian future.


14. Although there are a lot of definitions of what a great power is, we will use the term to refer to those states that can play a significant role in changing global balances of power. In this sense, a hypothetical increase in military aggressiveness by Peru might be worrying for its neighbors, but it will not affect the global balance of power. A hypothetical increase in aggressiveness by China could start modifying significantly the world’s balance of power, and thus we consider China a great power. John J. Mearsheimer’s definition is as follows: “[g]reat powers are determined largely on the basis of their relative military capability. To qualify as a great power, a state must have sufficient military assets to put up a serious fight in an all-out conventional war against the most powerful state in the world.” John J. Mearsheimer, The Tragedy of Great Power Politics (New York: W.W. Norton & Company, Inc., 2001), 5.


18. Nick Bostrom argues technology “is the sum total of instrumentally useful culturally-transmissible information (…) to include not only gadgets and machines but also techniques, processes, and institutions.” Thus, we can say that democracy is also a technological development. Nick Bostrom, “The Future of Humanity,” New Waves in Philosophy of Technology, eds. Jan-Kyrre Berg Olsen, Evan Selinger and Soren Riis (New York: Palgrave MacMillan, 2009), 189. From this definition, the anarchic international system is a technological development. It exists due to the inability of mankind to create a world government or a new type of international political economy which allows the international system to be based not on the maximization of power for hegemony but on a new humanistic vision of the world. So far, mankind does not have a concrete answer on how to resolve this.


33. Offensive Realism does not postulate that states never cooperate, but that survival and security are their main objectives, overriding any others. John J. Mearsheimer, Power and Fear in Great Power Politics, 192.
40. Stockholm International Peace Research Institute, “World Military Spending: Increases in the USA and Europe, Decreases in Oil-exporting Countries.”
50. A typical example briefly illustrated here is the development and use of the first nuclear bombs. Were the nuclear attacks on Hiroshima and Nagasaki necessary to end the war in the Pacific front? It is often argued that they were necessary to avoid masss of U.S. causailities in the Pacific, but more nuanced historical revisions tell us that by then most of Japan was razed to the ground and that most of the fighting on the Pacific front was being done by the Soviets; and even when the bombs were dropped, Japan did not surrender until the allies assured Japan that its emperor would not be humiliated. So, why drop these bombs? The war was ending, the weapons had already been developed, and the United States saw the opportunity to project its power and to send a clear signal to the world: the United States would be the new world power. For this it needed to go beyond traditional weapons. At Potsdam, Truman bragged to Stalin about a new type of weapon that could unleash unusual power. Marshal Georgi K. Zhukov wrote in his memoirs that that same day, Stalin ordered his scientists to rush the research on their atomic bomb. The Cold War had started.
58. Following OR, if the survival of a state is in question, it will override international norms. For example, after the attack on Pearl Harbor, the United States reignited submarine warfare, destroying Japan’s civilian naval fleet (a war crime), ending a decade of an international treaty banning submarine warfare that the United States had helped forge. When any great power is threatened or aims to achieve regional hegemony, it feels justified in breaking international laws, as when Germany broke the Molotov–Ribbentrop Pact (German–Soviet Non-aggression Pact) in World War II; the examples are numerous.

**BIBLIOGRAPHY**

Offensive Realism and the Insecure Structure of the International System


