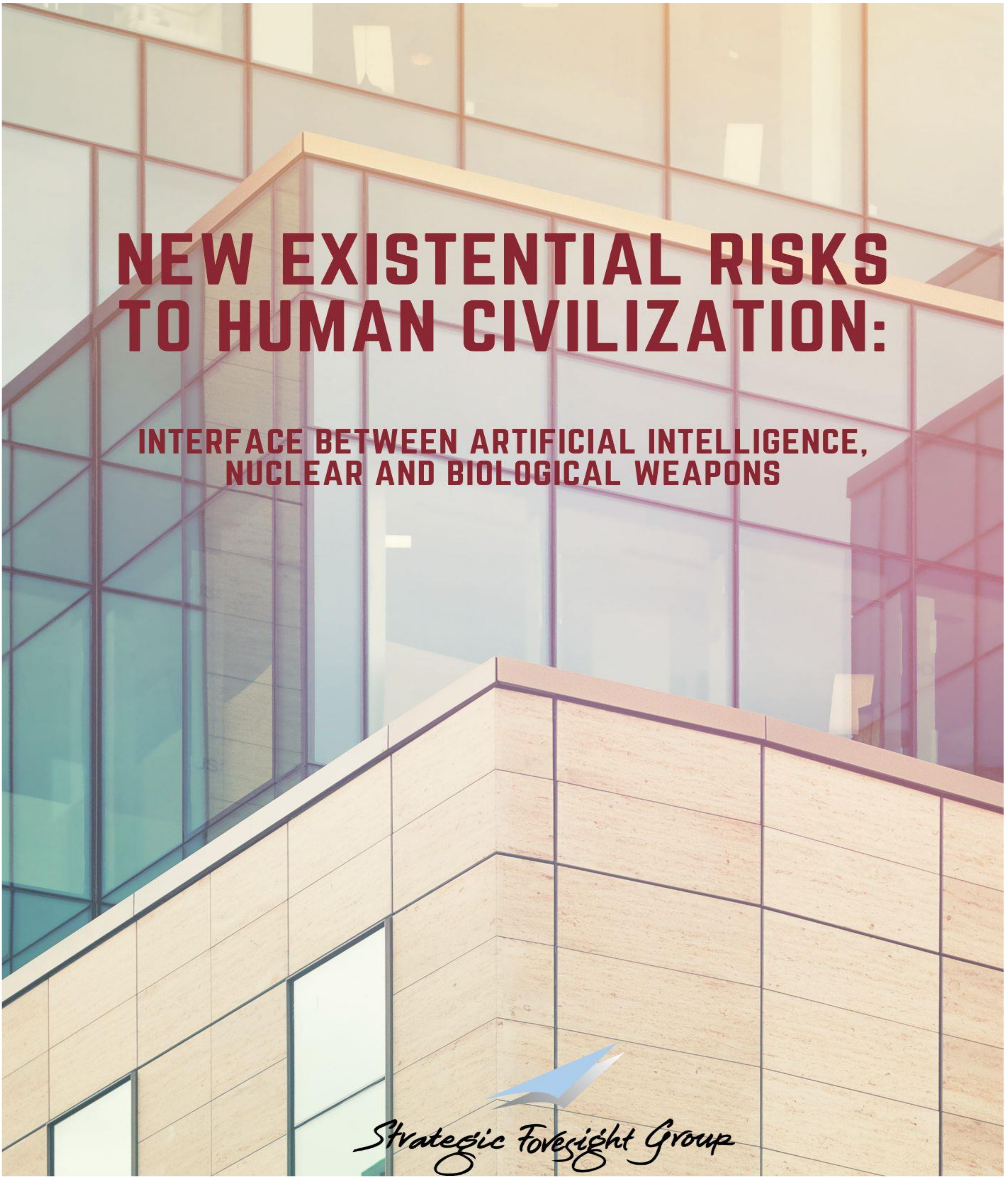
NEW EXISTENTIAL RISKS TO HUMAN CIVILIZATION:

INTERFACE BETWEEN ARTIFICIAL INTELLIGENCE, NUCLEAR AND BIOLOGICAL WEAPONS



Strategic Foresight Group

NEW EXISTENTIAL RISKS TO HUMAN CIVILIZATION:

INTERFACE BETWEEN ARTIFICIAL INTELLIGENCE, NUCLEAR AND BIOLOGICAL WEAPONS

Occasional Paper

Sundeep Waslekar, Thomas Greminger, Alexandra Matas, Ilmas Futehally

December 2022

Strategic Foresight Group

CONTENTS

KEY POINTS

- The growing interface between artificial intelligence, nuclear weapons and biological weapons poses existential risk to our species.

- While most of these technologies and their potential applications remain under development, the potential dangers posed by them would be so catastrophic that careful consideration of the implications of these developing risks would be advisable.

- The year 2022 began with hope as the 5 Permanent Members of the UN Security Council (P5) issued a joint pledge on 3 January 2022 to lower the risks of nuclear war. The US and Russia held 3 rounds of strategic stability talks in 2021. There was some speculation about bilateral strategic talks between the US and China.

- Since the Ukraine war beginning in February 2022, dialogue between the P5 has broken down at all levels. The visit of the US Congressional delegation to Taiwan in August 2022 has resulted in the suspension of military dialogue between the US and China. The NPT Review Conference in August 2022 resulted in a failure, without a consensus statement.

- While the channels of communication have been suspended between the P5 countries, the strategic arms race continues, further complicated by the potential interface of artificial intelligence (AI) and cyber technology with nuclear weapons and the interface between AI and biotechnology.

- There are accidental, inadvertent, third party and deliberate risks of a catastrophic nature arising from the interface between AI and nuclear weapons. These include hypersonic missiles that could chart their own trajectory, the manipulation of early warning systems, risks associated with unmanned under water vehicles, among several others.

- Experts disagree on which one of such risks are immediate, plausible in distant future, and only imagined in theory or science fiction.

- The US, Russia and China have varied perspectives depending on their comparative strengths. The US and its allies cooperate with the private sector to avail of additional AI related manpower resources. China has access to manpower in the government sector. Russia faces relative deficit of manpower as well as hardware. However, Russia and the US are ahead of China since they initiated strategic arms development several decades ago.

- The US, Russia and China have all made AI for military purpose a top priority in their future strategy and are accordingly building their physical capacities.

- There is a need to restore the dialogue between P5 countries at whatever level and in whatever form possible.

- There is a particular need to focus on the interface between AI and nuclear weapons in mapping and implementing risk reduction measures.

- Efforts should be made to keep the “human in the loop” and not hand over important nuclear related decisions to AI.

- Lethal autonomous weapons, drones and cyberweapons can emerge as a new form of delivery systems for weapons of mass destruction, and therefore human control on such weapons must be consciously increased.

- There is a need for inter-disciplinary experts' dialogues on AI related issues. There should be conversations involving weapons designers and other relevant scientists from the P5 countries to discuss risk reduction processes.

- The writers of AI algorithms from P5 countries should meet and challenge each other to find solutions for de-escalation and share knowledge.

- A charter of good governance in AI should cover the use of AI in command and control of nuclear weapons.

INTRODUCTION

We are pleased to present this paper to promote dialogue and discussion within the framework of the Normandy P5 Initiative on Global Security and Catastrophic Risks.

The Normandy P5 initiative is inspired by the Normandy Manifesto for World Peace, which was issued at Caen, Normandy on 4 June 2019 by six thought leaders, including four Nobel Peace Laureates, from different parts of the world. It calls for saving humankind from an apocalypse which might be caused by weapons of mass destruction including nuclear, biological, genomic, chemical, and lethal autonomous weapons, at a time of erosion of global values, rise of ultra-nationalism and the danger of gradual surrendering of human control on deadly weapons to machines. The initiative is convened by Strategic Foresight Group (SFG), Geneva Centre for Security Policy (GCSP), and Normandy for Peace Initiative of the Region Normandy.

The overall spirit of the initiative was explained by the heads of the three convening institutions in a blog article published by the GCSP and sections of the French print media.

“The greatest challenge facing us is to prevent these two cold wars (US-Russia and US-China) from sliding into a great war of human extinction. It is naive to believe that the future use of nuclear weapons will be limited to attacks using tactical weapons on a few cities. A future world war will involve the use of thousands of missiles each carrying multiple nuclear warheads. In a matter of hours large swathes of humankind will be annihilated, while those who survive the initial attacks will be even more unfortunate, dying a slow and painful death from radiation sickness and a nuclear winter. Preventing such a great war must therefore be our main priority. Among other things, this will require us to rebuild the global security architecture that has been increasingly fragmented by the two cold wars, and particularly by the war in Ukraine.

Any hope for reforming the global security architecture will depend on the political will of the P5 countries. It is an urgent need of our time to engage these countries in the discourse on the need to build a robust, resilient, and reliable global security architecture for this century. We could begin this discourse with discussions among experts and then gradually expand it to include states’ political representatives.”

The GCSP and SFG held consultations with P5 Disarmament Ambassadors based in Geneva in June and August 2022.

In September 2022, the three convening organizations brought together experts from US, France, UK, and China in a roundtable at Caen, Normandy to discuss risk reduction measures.

This paper has benefited from these diplomatic consultations as well as experts’ roundtable. However, the responsibility for content is that of the authors alone.

The paper underscores the urgent need for strategic arms risk reduction dialogue between the P5 countries at a time when the strategic communication between the West and Russia as well as between the US and China has been suspended. However, the unfortunate deterioration of the last months should not prevent us from creating a soft infrastructure of ideas for a reformed global security architecture. The alternative is to allow the cold wars between big powers to trigger an incident or accident that could end human civilization.

EVOLUTION OF STRATEGIC CONCEPTS

A. Nuclear Weapons

There are about 10,000 operational nuclear warheads out of the total warheads of 13,000 in the world. Even though this number represents a reduction from the stockpile of over 60,000 weapons in the 1980s, it is more than enough to cause catastrophic consequences that would alter life for every inhabitant of the planet. There are 2800 nuclear weapons on hair-trigger alert, ready to be launched within 5-10 minutes of an executive decision by the leader of the United States or the Russian Federation. UN Secretary General Antonio Guterres said in his address to the opening session of the NPT Review Conference in August 2022 that humanity was one miscalculation away from nuclear annihilation. In the 2020s, with the induction of low yield weapons, nuclear weapons may be seen as becoming more "usable," as they will be presented as small warheads with a yield less than the bombs dropped on Hiroshima and Nagasaki in 1945. The debate on low yield warheads does not dwell much on the fact that many of them are earth penetrating weapons, capable of 20 times the blast yield of their declared power, if detonated beneath the surface.

There is a dialogue mechanism for strategic risk reduction between the heads of disarmament divisions of the P5 countries. This mechanism produced an extraordinary P5 consensus statement, also known as the joint pledge, on 3 January 2022 to affirm "the avoidance of war between Nuclear-Weapon States and the reduction of strategic risks as our foremost responsibilities." The joint pledge reiterated the Reagan-Gorbachev statement of 1985 saying: "We affirm that a nuclear war cannot be won and must never be fought." It also reaffirmed commitment to the NPT Article 6 obligation towards general and complete disarmament. However, the statement offset its apparent spirit and commitments by introducing an emphasis on deterrence in the text and failed to commit to any concrete measures such as adherence to no first use or shifting all weapons away from hair trigger alert. The doctrine of deterrence is not in harmony with the doctrine of no first use.

The spirit of the 3 January 2022 was further diluted when only France, UK and USA issued a similar statement on the avoidance of a nuclear war at the opening of the NPT Review Conference on 1 August 2022, without China and Russia. Their statement strongly condemned Russia for its invasion of Ukraine. On the other hand, China strongly condemned the United States in its statement to the conference. The P5 are thus split. However, in consultations with a delegation of the Geneva Centre for Security Policy and the Strategic Foresight Group, the P5 disarmament ambassadors have individually confirmed their commitment to the 3 January 2022 statement on the avoidance of nuclear wars. The split within the P5 and the differences of opinion between nuclear and non-nuclear weapons states resulted in the failure of the NPT review conference to arrive at any consensus at the end of August 2022. The Stockholm Initiative for Nuclear Disarmament has proposed a set of nuclear risk reduction measures, but the initiative does not involve any nuclear weapon states.

It is generally believed that a nuclear war has not taken place because of deterrence. Empirical evidence, however, is ambiguous. Although there has not been a nuclear war, there have been many incidents of close calls when the world came close to a nuclear war which was avoided because of courageous decisions taken by certain officers or a timely comprehension of the accidental nature of the risk. The Future of Life Institute has published a list of such close calls on their website. Evidence proves that the avoidance of nuclear war cannot be taken for granted.

The growing application of artificial intelligence (AI) has further increased the existential risk of a nuclear war by accident, incident, or intent. The application of AI in NC3 is particularly ambiguous and potentially dangerous. At the same time, the application of AI in biotechnology has given rise to the risk of a new generation of biological weapons of mass destruction.

> **If AI is used in nuclear weapons, their delivery systems, and new types of biological weapons, the human race could become extinct, along with other life forms, in a global war. Thus, the use of AI in nuclear and biological fields may eliminate the only known intelligent civilization in the universe and nobody but we ourselves - humans and our leaders, blinded by nationalist ambitions - would be responsible for our universal death.**

B. Artificial Intelligence

Artificial intelligence is a term coined in the 1950s and refers to the abilities of machines to mimic human behaviour associated with human intelligence such as visual recognition, natural languages processing or learning. It is composed of several techniques but those most of concerns to nuclear strategy are in the field of machine learning and automation.[1] Indeed, the overarching risk is that the combination of machine learning and automation embodied in AI generates extraordinary **speed**, thus reducing decision making time for humans receiving such input and making them susceptible to mistakes. A direct consequence of the higher speeds is compression of timeframes available to decision makers. The margins for de-escalation are invariably reduced, with the possibility of strategic destabilization. The second overarching risk is that the tools and weapons produced with AI mostly operate in **stealth**, making it difficult to detect them - leading to various scenarios including some based on misinterpretation of signals. AI depends on access to massive data. In the case of both nuclear and biological attacks, there is a paucity of data and therefore pattern recognition may have to take place with computer simulations rather than credible data. This can lead to mistakes. The risk posed by AI in nuclear command, control, and communication (NC3) is ambiguous. Even if AI is not directly used in command and control, automation and speed which have been increasing in command and control can generate risks. The use of AI and cybertechnology in communication and ISR (intelligence, surveillance, reconnaissance) is already growing.

C. Biotechnology

The discussion on artificial intelligence and nuclear weapons generally does not extend to other weapons of mass destruction. However, it is necessary to examine the impact of artificial intelligence on biotechnology, as some emerging research has dual use implications. It can be mainly seen from three angles. The first is gene-editing with CRISPER CAS 9 which enables scientists to modify genes before the birth of a child. If it were to be used for gene-line manipulation, an entirely new type of population could be created. The second is synthetic biology which enables scientists to produce life in a laboratory. Thus, a new kind of pathogen can be created with artificial intelligence, with properties either beneficial or dangerous for humankind. The third is creating a chimera which enables scientists to blend human genes with the genes of other species, giving birth to a new kind of life. Scientists are

[1] Rickli, Jean-Marc (2019). "The Destabilizing Prospects of Artificial Intelligence for Nuclear Strategy, Deterrence and Stability," in Boulanin, Vincent (ed.). *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk: European Perspectives.* Stockholm: Stockholm International Peace Research Institute, Volume I, pp. 91-98, https://www.sipri.org/sites/default/files/2019-05/sipri1905-ai-strategic-stability-nuclear-risk.pdf

experimenting with these three types of innovation with a view to service humankind by eliminating certain types of disease and creating sources of organs which are currently in short supply. However, there is a risk of some of these experiments going wrong and resulting in a generation of biological weapons of a Frankenstein, either by intent or accident.

TYPES OF RISKS

Researchers have outlined several catastrophic risk scenarios resulting from the use of AI in nuclear weapons and biotechnology. They can be categorised as 1) accidental 2) inadvertent 3) third party exploitation 4) deliberate risks.

The most obvious pathway to accidental risks in the use of nuclear weapons is the use of machine learning to facilitate autonomous early warning systems and intelligence, surveillance, and reconnaissance (ISR). In this case signals misinterpretation or data poisoning can occur, which can lead to errors that are not only unpredictable but also undetectable, escalating the situation to the imminent launch of missiles with nuclear payloads.

Furthermore, accidents can also occur due to overreliance on AI, also known as automation bias, wherein the humans in charge of decision-making end up putting more credence in the machine rather than in their own judgment and experience. The growing automation of tasks in analysis and decision-making could lead to a flash crash, resulting in unintended consequences. Proponents of automation, on the other hand, have argued that increasing the input from AI in the decision cycle can reduce the possibility of human error – a significant factor in most of history's close calls involving nuclear weapons. In the case of synthetic biology, an accident can happen by the unintended release of a dangerous pathogen outside the laboratory and its spread across boundaries.

The pathway to inadvertent risks turns operational when certain actions involving AI in a conventional war are misinterpreted by one of the parties to the conflict, leading to the use of nuclear weapons. Example: In the case of autonomous UAVs (Unmanned Aerial Vehicles) and UUV (Unmanned Underwater Vehicles) being deployed, several escalatory scenarios can develop. Even though deployed for remote sensing operations, there is a risk that an adversary might assume that they have been deployed for either conventional or even nuclear attack. In a further escalation, the deployed vehicle carrying conventional arms can be mistaken to be carrying nuclear weapons. Faced with a possible nuclear attack and thus a risk to the survival of its nuclear deterrent (second strike capacity), the adversary may opt for a pre-emptive strike. An Unmanned vehicle can also cause confusion wherein a state may assume it belongs to an adversary and, depending on geopolitical situation and level of acrimony and mistrust, nuclear weapons might at least be put on high alert. The adversary state, not being responsible for the unmanned vehicle could perceive this as a provocation.

The pathway for third-party exploitation would involve a third party, whether a state or a terrorist group, creating deepfakes that are very hard to detect. Non state actors that have vested interests in specific theatres can take advantage of the animosity and hostile relations between two states and foment a situation using deepfakes that can lead to brinkmanship. This could be done, for example, by releasing a fake video demonstrating a potential nuclear attack from one state, which could cause the other state to adopt extreme measures including pre-emptive strikes. This would be a deliberate attack/escalation, but it would be initiated by false information relayed by third parties. Moreover, if one of the states has already signalled towards its use of AI as a force multiplier and routinely flaunts its conventional forces, chances of the adversary believing the deepfake are higher. The most dangerous case of the third-party exploitation would be data poisoning in the early warning system of one of the two states, motivating it to launch a nuclear attack on its adversary.

Deliberate escalation by one state could evolve from either accidental or inadvertent risks from another. A pre-emptive first strike based on wrong information from AI early warning systems or ISR can result into deliberate escalation from the opponent state.

Deliberate escalation is also a likely scenario in case of asymmetry created by AI leaps favouring one state. To offset this advantage the adversary might engage in brinkmanship or launch an actual conventional attack. In extreme cases if the state fears the survivability of its nuclear deterrent capability, it might also opt for a pre-emptive first strike.

Since deliberate escalation by intent may lead to existential threat to human civilization, experts rule it out.

> **It is generally believed that no leader would deliberately want to put an end to human race. Historical evidence, however, is to the contrary. Not only leaders, but also the democratic opinion of populations, have many times led to wars causing worldwide carnage.**

The pathway to deliberate risks would involve eliminating the second-strike capacity of the adversary, thus dismantling the basis of deterrence. Without a second-strike capacity, there might be temptation to launch a massive first strike. AI technologies could enable sharpened tracking and targeting of ICBMs enclosed in silos, as well as submarines and mobile stealth vehicles, making it possible to use conventional warfare to launch attacks on them. Such capabilities would be especially destabilizing because decision-makers could threaten to employ conventional weapons much more plausibly than any kind of nuclear attack. A conventional threat would place the adversary under enormous pressure during a crisis, which could force it to capitulate or spiral into nuclear war. Such a deliberate escalation could happen if the adversary feels the need to use its nuclear weapons before its striking capacity is obliterated or after an unsuccessful strike on its deterrent forces.

The risks could take many specific forms, in terms of specific technologies and methods that are used by state parties involved in AI-driven nuclear or biological arms race.

- Hypersonic missiles which are potentially more difficult to track than ICBMs as they travel through the atmosphere and have autonomous manoeuvrability capacity, carrying nuclear payload
- Autonomous unmanned underwater vehicles with nuclear torpedoes capable of autonomously launching nuclear attacks from a submerged position in the sea
- Swarms of drones and other technologies for the defeat of mobile and underwater nuclear delivery systems, as well as missiles in silos, to neutralize the nuclear second-strike capacity of the enemy
- Lethal autonomous weapon systems potentially carrying nuclear payload in future (though currently most LAWS are not seen carrying nuclear payload)
- AI in nuclear command, control, and communication (NC3) will increase the pace of operations, causing pressure to reduce decision making timelines, which can lead to existential risk

- Genuine data misinterpretation in autonomous early warning systems and intelligence, surveillance, and reconnaissance (ISR) systems
- Deliberate data poisoning in autonomous early warning systems and intelligence, surveillance, and reconnaissance (ISR) systems
- Misinterpretation of the movement of unmanned air or water vehicles used for remote sensing operations as an imminent nuclear attack
- Automated decision-support tools to provide early warning of a pre-emptive nuclear attack being inherently risky as there are not enough real-life examples to adequately feed to the system that can indicate what an actual nuclear attack would look like
- Cyber-attacks on nuclear facilities of the enemy
- Synthetic biology producing new killer pathogens using pattern recognition in genomic data
- Gene editing leading to genetic manipulation of existing benign pathogens into harmful pathogens
- Creation of biological agents to target specific populations such a race, an enemy society, or an ethnic minority.

The forms of risks described above reflect technological developments known as of December 2022. There would be some technologies which are being developed but not known through open sources at this time. New technologies will appear on the horizon in future.

EXPERTS' DISAGREEMENTS

Rand Corporation says in a report that experts in the fields of AI and nuclear weapons and the intersection between the two do not agree over the extent to which AI incorporated nuclear systems pose risks to global security. Experts with differing views in these regards have been classified as 'complacents', 'alarmists' and 'subversionists'.

By and large, as the taxonomy suggests, the 'complacents' are of the belief that the fear of additional risks that machine learning and autonomy can pose in strategic instability is unfounded as the risks are negligible. They are of the view that the chance of heavy incorporation of AI in the nuclear weapon systems is slim to be begin with, and in case of such an integration, the premise that AI can considerably destabilise nuclear balance is unlikely. The 'alarmists' have an exactly opposite view and are summarily against use of machine learning and autonomy to enhance or enable nuclear weapon systems, especially in command-and-control systems. The 'subversionists' hold the view that regardless of development and capabilities of AI and the degree of its integration in weapon systems, there is a substantial risk of conflict intensification on account of susceptibility of AI to cyber-attacks like hacking, spoofing, data poisoning etc. They posit that an adversary could try to mislead, swindle, and deliberately cause misperception which can consequently also beget destabilisation and escalation making it a risky affair. [2]

[2] Edward Geist and Andrew Lohn, How Might the Artificial Intelligence Affect the Risk of Nuclear War, Rand Corporation, 2018, https://www.rand.org/content/dam/rand/pubs/perspectives/PE200/PE296/RAND_PE296.pdf

To understand which viewpoint is most reasonable, it is important also to understand how major nuclear weapons states are using AI and its applications for civilian and military purposes. This can help researchers comprehend whether complacency, alarmism, subversionism or a mix of the three is the most appropriate approach to understanding the ontology of nuclear doctrines in an epoch defined by advancements in robotics and AI.

COMPARATIVE PERSPECTIVES

Among the P5 nuclear powers, US, UK, and France are relatively open in articulating their positions and releasing some information. The United States government has been forthcoming about publishing policy documents delineating its intentions to harness machine learning, like the American AI Initiative of February 2019. The publication of the US Department of Defence's (DoD) Defence Innovation Initiative in 2014, also known as the Third Offset Strategy, was the clearest evidence that the USA once again considered AI to be a game-changing technology in warfare and national security. The Department of Defence too has created Defence Innovation Unit (DIU) and Joint Artificial Intelligence Centre (JAIC) to foster research and learning as well as development of military applications of AI. This is clearly indicative of US's acknowledgement of the relative advantage it has and its intention to explore it aggressively. US DoD has been looking for collaborations with the private sector to further this agenda. This is mainly because USA, unlike China, does not have the advantage that centralisation offers. DoD cannot match the remuneration packages that IT giants can offer and thus, collaboration and joint ventures is a clever way for it to stay ahead in the game. Similarly, the UK is also seeking partnership with the private sector in AI development. The National Security Investment Fund invests in venture capital companies that back high-tech start-ups. Even MI6 - the UK intelligence agency - is investigating prospects for cooperation with technology companies to counter cyber security threats from the adversary states.

Various statements of US, UK and French officials indicate that they are most concerned about China's capabilities in artificial intelligence, cyber surveillance of the world and hypersonic missiles. They seem to be relatively less concerned about similar capabilities of Russia.

With respect to its stance on LAWS, the US government has remained steadfastly against a preventive ban. However, at the same time the military has also emphasized that the reason for this is because of the humanitarian benefits that use of AI can bring about by helping with precision. US has maintained that autonomous weapons by themselves are not the problem, if used and operated ethically with adequate human control. They have also stressed, much like Russia, that existing international conventions and regulations are sufficient.

Russia has been showing particular interest in AI development in recent years. In 2017, Russian President Vladimir Putin said, "Artificial intelligence is not only the future of Russia, it is the future of all mankind. There are enormous opportunities and threats that are difficult to predict today. The one who becomes a leader in this sphere will be the ruler of the world." This statement of President Putin indicates Russia's willingness to acquire superior AI capabilities as well as provides a glimpse into its probable vision. In 2014, the National Defence Operations Centre was inaugurated for collection of information, analysing data based on prevalent geopolitical equations and to maintain a centralised control of Armed Forces.[3] Using open-source material and data analytics, automated recommendations about action to be undertaken are made.

Though Russia has always maintained a kind of bluster proclaiming that the US poses no threat to its nuclear triad, recent advancement in US capacities is bound to raise anxiety. It also serves as a good incentive for Russia to attempt to strengthen its nuclear deterrence. Already Russia has been working on modernising and innovating its weapon systems, for instance, its development of hypersonic missiles and unmanned underwater nuclear weapons like Poseidon.

Currently though, Russia's development of AI/ML is in its early stage as per several indicators. Many factors play a role in this. One of the main shortfalls is the dearth of human resources. Severely lacking in AI scholars, researchers, and scientists, unlike US and China, Russia is likely to lag. Moreover, unlike China (and to an extent US) massive amount of data required for rigorous machine learning is not accessible to Russia. In addition to this, some researchers believe that the Russian Federation relies on foreign sources for essential hardware required to develop AI.

Much like other nuclear capable states, Russia is of the opinion that adequate leeway for responsible use of AI should be the goal. They have also been emphasising on properly defining the scope of LAWS. Russia wants LAWS to be defined as 'an unmanned piece of technical equipment that is not a munition and is designed to perform military and support tasks under remote control by an operator, autonomously or using the combination of these methods'.[4] Such taxonomy may offer more confusion than clarity. But the bottom line is that Russia opposes a ban on LAWS.

In China, People's Liberation Army is already making headway in research, development, and experimentation with regards to artificial intelligence and its various applications. The New Generation Artificial Development Plan released in July 2017 is a good indicator of China's resolve to be a leader in the field of AI and not just follow the footsteps or replicate technology originating in the US. Xi Jinping's statement, "Under situation of increasing fierce international military competition only innovators win", is also testimony to China's attempt at being a leading nation for AI development. In October 2018, the Beijing Institute of Technology (BIT), one of China's top weapon research institutes, launched a four-year 'experimental programme for intelligent weapons systems' at the headquarters of NORINCO, a military industrial company. The PLA rocket force has upgraded its nuclear ammo by including Multiple Independent Re-entry Vehicles (MIRVs). Considering the US 2022 Nuclear Posture Review explicitly mentioning China as a nuclear threat that needs to be countered, China is likely to modernise its nuclear force for the fear of its second-strike capability being threatened.

China is considered a hub of data and is set to soon be in possession of about 20 percent of global data. This opens up a fertile ground for machine learning. The fusing of military and civilian capacities in China shows that as it makes advancements in machine learning, there is a risk of "intelligising" of weaponry in subtle ways.

[3] By Heather A. Conley, Vladimir Orlov, Gen. Evgeny Buzhinsky, Cyrus Newlin, Sergey Semenov, and Roksana Gabidullina, The Future of US-Russian Arms Control, Centre for Strategic and International Studies, Washington DC, March 2021, https://www.csis.org/analysis/future-us-russian-arms-control-principles-engagement-and-new-approaches

[4] Lora Saalmaan, The Impact of AI on Nuclear Deterrence: China, Russia and the United States, East West Centre, https://www.eastwestcenter.org/news-center/east-west-wire/the-impact-ai-nuclear-deterrence-china-russia-and-the-united-states

China has an ambivalent attitude towards LAWS, opposing their use but not objecting to research, production, and development. Overall, utmost secrecy and centralised control of military affairs, makes it difficult to assess Chinese perspectives on the interface between AI and nuclear weapons for an outside observer.

A comparison between the United States/allies, Russia and China reveals that the US and Russia face the deficit of human resources and big data, as compared to China. However, the Americans can meet this deficit through collaboration with the private sector, an option Russia does not have. The Chinese are complementing their human resources by acquiring talent in countries like Ukraine ostensibly for civilian use of AI, but not ruling out their dual use. Despite China having some structural advantages, it is a late comer in nuclear, biological and AI race and the US and Russia have advantages accrued by decades of experience in these spheres.

The US intends to build a missile shield to ensure the survival of its nuclear arsenal in case of a limited attack from an adversary such as North Korea or an accidental launch, while using AI to ensure a second-strike capacity against Russia and/or China to deter any attempt at a pre-emptive first strike.

> **Russia and China run the risk of their second-strike capability being terminated, and therefore the theory of deterrence being nullified, it will not be surprising if they prepare for an effective first strike and develop technology and strategies to bypass the American missile shield.**

RISK REDUCTION

On 7 December 2021, the P5 nuclear powers submitted a joint working paper to the NPT Review Conference on strategic risk reduction. It emphasised dialogue between the P5 states on reducing the risk of the use of nuclear weapons and reducing the risk of armed conflict between nuclear powers. To this end, it emphasised promoting strategic stability and predictability. It recommended the strengthening of existing bilateral dialogue channels, formalised communication between diplomatic and military channels to clarify concepts and doctrines, the use of communication tools such as hotlines and the notification of missile launches and resolving conflicts arising out of misinterpretation of the adversary’s intentions. However, since the beginning of the war in Ukraine in the last week of February 2022, most dialogue and communication channels within the P5 countries have broken down. There is no communication at the diplomatic or military level. The official P5 dialogue mechanism on nuclear risk reduction, non-proliferation and disarmament has been suspended. There has not been any collective interaction, formal or informal, between the disarmament ambassadors of the P5 countries accredited to the UN Conference on Disarmament in Geneva since February 2022.

In the absence of an official risk reduction dialogue process, researchers have identified several ideas to reduce risk from the use of AI in military affairs in conjunction with nuclear weapons at present or potentially in future.

There is a movement demanding a total ban on Lethal Autonomous Weapon Systems (LAWS), which is opposed by the P5 countries. Currently, LAWS are not associated with nuclear weapons but that is not a guarantee that they could not be used to deliver nuclear payload in future. The Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects as amended on 21 December 2001, is usually referred to as the Convention on Certain Conventional Weapons or CCW. The purpose of the CCW is to ban or restrict the use of specific types of weapons that are considered to cause unnecessary or unjustifiable suffering to combatants or to affect civilians indiscriminately.[5] A group of governmental experts (GGE) was formed a few years ago and last met in July 2022 to discuss the questions related to emerging technologies in the area of lethal autonomous weapons systems (LAWS). It is necessary to assess the observations made by GGE in their meetings in 2022 and reflect on further actions.

It is necessary to promote dialogue and cooperation at the multilateral level on the use of AI in conjunction with nuclear weapons. This may include issuing specific and inflexible limits on the use of AI technology in nuclear weapon systems. An unambiguous declaration of sorts with checks and balances that define defaulters can go a long way setting an effective, verifiable standard. A higher level of multilateral cooperation may go beyond setting norms and enable technology sharing, transparency and confidence building measures.

The US began strategic stability talks with Russia on nuclear weapons in 2021, though they were suspended in the aftermath of the Ukraine war in early 2022. There is a need for strategic stability talks between the US and China. These talks would be concrete risk reduction measures whenever they take place, though China announced the suspension of all military dialogue with the US following the visits of US Congressional delegations to Taiwan in August 2022. It would be useful to examine if the (revived) US-Russia and (potential) US-China parallel talks should be complemented by trilateral US-Russia-China or P5 strategic stability talks. The joint pledge issued by the P5 countries on 3 January 2022 commits to risk reduction dialogue between the P5 countries.

International Atomic Energy Agency (IAEA) could perhaps play a crucial role in the evaluation of the state of AI technology in nuclear weapons infrastructure, with its verification and monitoring capabilities. By enhancing its watchdog ambit, the IAEA can make use of its existing credibility and authority, to bring attention to the insidiousness of military applications of AI in nuclear weapon systems.

Civil society and tech workers can play a critical role in reducing the use of AI for weapons of mass destruction. The Future of Life Institute has successfully managed to get more than 2400 signatories to the pledge against helping militaries and states in building LAWS. This move by the community of scientists working for companies of the calibre of SpaceX, DeepMind, show a commitment towards ethics over what may sometimes be purported as national imperative.[6] Collaborative creation of knowledge systems that clearly delineate ethical use of technologies that can be catastrophic, especially in the context of nuclear and biological weapons, can help create spaces for advocacy of humane use of AI technology.

The 1972 Biological and Toxin Weapons Convention (BTWC) is the central governance instrument for biological arms control. It is severely lacking in the coverage of the applications of AI and robotics. The governance frameworks of BTWC have not yet explored in their entirety how to include convergence of emerging technologies. The BTWC regime can be reformed and a dedicated BTWC Scientific Advisory Board can be established to examine the issue of convergence between potential biological weapons, artificial intelligence, synthetic biology, and other technologies.

[5] https://www.un.org/disarmament/the-convention-on-certain-conventional-weapons/

[6] https://www.theguardian.com/science/2018/jul/18/thousands-of-scientists-pledge-not-to-help-build-killer-ai-robots

Martin Dahinden has argued in a paper published by ICT4Peace that arms control approach should be applied to cyberweapons. An ambitious overarching global agreement is not possible, due largely to challenges involving attribution and verification. However, specific measures should be explored including no first use of cyber weapons, agreement on not to target civilian targets, as well as not to target nuclear weapons infrastructure and command and control facilities (as it can lead to escalation), promotion of dialogue between protagonists, and cooperation by state parties against cyber-attacks launched by terrorist groups and other non-state parties.[7]

Other risk reduction measures may include declaratory initiatives by nuclear weapon states to proclaim their long-term intent to prevent wars of human extinction due to convergence between artificial intelligence, other technologies, and weapons of mass destruction. The declaratory initiatives can be gradually transformed into legal binding agreements and treaties.

A group of experts meeting in a roundtable at Caen, Normandy, in September 2022, suggested specific strategic risk reduction measures, in addition to the restoration of dialogue between all P5 countries mentioned earlier. These include:

1. The Reagon-Gorbachev Declaration on non-use of nuclear war should be reiterated time and again.
2. The renewal of New START Treaty by 2026 is necessary. It is also necessary to begin thinking about arms control beyond 2026. It will be useful to think about the revived New START Treaty further reducing the number of weapons allowed to be deployed. If the New START Treaty and particularly verification measures prove difficult to seek the approval of the Senate in the US and Duma in Russia, then another form of working agreement rather than a treaty can be negotiated.
3. Since the New START treaty is between the US and Russia, it is necessary to have a separate strategic arms control regime between US and China.
4. We need to urge leaders to look at the arms control regime, including the treaties that have been abandoned.
5. The P5 states can take unilateral measures without always seeking reciprocity.
6. The gap between actual and perceived capabilities increases risks. We need measures to reduce such gaps.
7. The P5 countries should negotiate phasing out of low yield nuclear weapons.

The experts' roundtable suggested some measures which are specific to the interface between AI and nuclear weapons. These include:

1. The dialogue mechanism which currently exists among the P5 members, but has been dysfunctional since February 2022, does not include the interface between AI, cybertechnology and nuclear weapons. A priority should be given to examining these linkages and identifying norms for their regulation in the official P5 deliberations. Experts should urge the officials to place such linkages on the intergovernmental agenda with a sense of urgency.

[7]

Martin Dahinden, How Can Arms Control and Disarmament Contribute to a Secure Cyberspace, ICT4Peace, Geneva, April 2022

2. Efforts should be made to keep “human in the loop” and not hand over important nuclear related decisions to AI.
3. LAWS, drones and cyberweapons can emerge as a new form of WMD delivery systems and therefore human control on such weapons must be consciously increased.
4. Some of the lethal autonomous weapon systems (LAWS), particularly unpredictable and anti-personnel AI weapons should be completely banned, while others should be highly regulated.
5. We need more transparency with regards to LAWS and hypersonic missiles.
6. We must address dual use systems, and clearly denominate what is a nuclear system in ambiguous situations.
7. It is necessary for scientists and officials to explore how to increase decision-making time in conflict management with a clear policy of reducing reliance on automation. We need a better de-escalation system in conflict management. We need to reduce systems on high alert and the number of committed land-based ICBMs.
8. There is a need for inter-disciplinary experts on AI related issues. There should be input from weapons designers and other relevant scientists to the risk reduction processes.
9. The writers of AI algorithms should meet and challenge each other to find solutions for de-escalation and share knowledge.
10. A charter of good governance in AI should cover the use of AI in command and control of nuclear weapons.
11. We need to develop communication systems to prevent data poisoning, which is in the interest of all P5 countries.
12. There is a need to share datasets from experiments made in the 1960s and 1970s to train algorithms to create scenarios for the future.
13. Automation should be kept away from nuclear command, control, and communication (NC3).
14. We need to create physical barriers in hair trigger alert in order to prevent the launching of nuclear attacks in haste and without due negotiations. We need better de-escalation systems.
15. Codes of Conduct on responsible behaviour in cyber-space and outer space are required.

CONCLUSIONS

First, there is ambiguity about the nature of risk posed by technologies involving in AI/nuclear weapons interface. It is necessary to identify which technologies pose imminent threat, which ones pose risks in future and which ones are of merely speculative nature.

Second, there is a complete breakdown of communication between P5 countries on nuclear risk reduction and disarmament issues with the suspension of the official dialogue mechanism as well as the informal interaction between P5 disarmament ambassadors since February 2022. It is important to open dialogue at formal or informal, official, or expert level, whatever might be feasible, as soon as possible.

Third, the NPT Review Conference failed to produce any consensus statement. There is no visible progress on talks on the renewal of the New Start Treaty beyond 2026. The Treaty on the Prohibition of Nuclear Weapons has not been accepted by any of the nuclear weapons states or their allies. The nuclear arms control regime is in tatters which is very dangerous. It is necessary to revive it.

Fourth, the P5 countries appreciate the role of artificial intelligence in increasing the unpredictability of the weapons of mass destruction but they did not have plans even before suspension of the dialogue mechanism in February 2022 to address this dimension. The P5 dialogue mechanism have reached an agreement on a common glossary and have identified nuclear doctrines and strategic risk reduction as the priorities to be addressed. It is necessary to add the risks associated with artificial intelligence on the P5 agenda.

Fifth, it is in nobody’s interest that accidental, inadvertent, and third-party manipulation risks result in a war of human extinction. It is necessary to explore risk reduction measures specifically aimed at these three types of risks. These measures should be wide ranging and include the interface between AI and nuclear weapons, cybertechnology and biotechnology.

REFERENCES

Normandy Manifesto for World Peace

https://normandiepourlapaix.fr/en/manifeste-pour-la-paix

Joint Working Paper on Strategic Risk Rection submitted by the P5 to the NPT Review Conference

NPT/CONF.2020/WP.33 dated 7 December 2021

Joint Statement by P5 countries on Preventing Nuclear War

https://www.whitehouse.gov/briefing-room/statements-releases/2022/01/03/p5-statement-on-preventing-nuclear-war-and-avoiding-arms-races/

ACKNOWLEDGEMENTS

A delegation of the Geneva Centre for Security Policy (GCSP) and the Strategic Foresight Group (SFG) held consultations with the Disarmament Ambassadors of the Permanent Members of the UN Security Council (P5 countries) in June and August 2022 within the framework of the 'Normandy P5 Initiative on Global Security and Catastrophic Risks'. The Geneva-based P5 diplomats included the following:

- Li Song, Disarmament Ambassador of People's Republic of China
- Yann Hwang, Disarmament Ambassador of France
- Andrey Belousov, Disarmament Ambassador of Russia
- Aidan Liddle, Disarmament Ambassador of UK
- Aud-Frances McKernan, Acting Permanent Representative of the United States to the Conference on Disarmament.

An Experts' Roundtable on Global Security and Catastrophic Risks was held at Caen, Normandy, on 22 September 2022 with the following participants:

- Ambassador Thomas Greminger, Director of Geneva Centre for Security Policy, former Secretary General of the Organization for Security and Co-operation in Europe (OSCE), co-chair
- Dr Sundeep Waslekar, President of Strategic Foresight Group, co-chair
- Mr François Xavier Priollaud, Vice President of Normandy Regional Government and Mayor of Louviers (Normandy Region), co-chair
- Mr Paul Ingram, Centre for the Study of Existential Risk, Academic Programme Manager and Senior Research Associate, Cambridge University, UK
- Dr Rebecca Johnson, AIDD Director (Acronym Inst. for Disarmament Diplomacy), ICAN Steering Group (2017 Nobel Peace), CND Vice President, UK
- Dr Dave Webb, Vice President, International Peace Bureau, chair of the UK Campaign for Nuclear Disarmament (CND), the Convenor of the Global Network Against Weapons and Nuclear Power in Space. Emeritus Professor of Peace and Conflict Studies at Leeds Beckett University (previously Leeds Metropolitan University), UK
- Dr Sebastian Brixey-Williams, Director, BASIC, Gender Champion in Nuclear Policy, an ELN Mentor, and a member of the Younger Generation Leadership Network on Euro-Atlantic Security. N Square Fellow working to improve the sustainability of the nuclear policy field (2020-2021), UK
- Ms Lynn Rusten, Vice President, Global Nuclear Policy Program at the Nuclear Threat Initiative. Former senior director for arms control and non-proliferation on the White House National Security Council staff, Department of State including chief of staff for the Bureau of International Security and Non-proliferation and senior advisor in the Bureau of Arms Control, Verification and Compliance (AVC), USA

- Prof Anthony Aguirre, Co-founder of the Future of Life Institute, Professor of Physics at the University of California, Santa Cruz, holding the Faggin Presidential Chair for the Physics of Information, Associate Scientific Director of the Foundational Questions Institute, USA
- Dr Nicole Gnesotto, Professor at Conservatoire national des arts et métiers, holder of the European Chair, Vice-President of the Institut Jacques Delor. First director of the European Union Institute for Security Studies. Former deputy head of the French Foreign Ministry's Centre d'Analyse et de Prevision, France
- Dr Benjamin Hautecouverture, Senior Research Fellow for arms control, non-proliferation and disarmament issues at the Fondation pour la Recherche Stratégique (FRS), France
- Dr Héloïse Fayet, Research Fellow at Ifri's Security Studies Center, focus on proliferation and dissuasion issues, military forces' capacities analysis and strategic anticipation. Formerly with French Ministry of Defense, France
- Prof Li Bin, Professor of international relations at Tsinghua University. Former director of the arms control division at the Institute of Applied Physics and Computational Mathematics, executive director of the Program for Science and National Security Studies, China
- Dr Qi Haotian, Assistant Professor at the School of International Studies of Peking University, Secretary General of the Institute for Global Cooperation and Understanding (iGCU) of Peking University, China
- Dr. Tianjiao Jiang, Associate Professor at Fudan Development Institute. Assistant director at Center for BRICS Studies and Cyberspace International Governance Research Institute, Fudan University, China
- Ms Adriane Bajon, Normandy for Peace Initiative, Region Normandy, France
- Ms Alexandra Matas, Geneva Centre for Security Policy, Switzerland
- Ms Ilmas Futehally, Strategic Foresight Group, India

We would like to thank the Future of Life Institute and the Silicon Valley Community Foundation for their support to the Normandy P5 Initiative which made this paper possible. We would like to thank the participants in the P5 Ambassadors Consultations as well as the experts' roundtable held on Global Security and Catastrophic Risks held at Caen, Normandy, on 22 September 2022. We would like to thank Dr Jean-Marc Rickli of GCSP for reviewing an earlier version of this paper. While we are grateful to these interlocutors, we would like to clarify that neither Future of Life Institute/Silicon Valley Community Foundation nor the participants in various interactions are responsible for the contents of this paper. The authors have full responsibility and moral right to the contents.