United Nations Security Council

Topic: Taking precautions against the possible risks of rapidly evolving artificial intelligence technology

Doga Sari & Beril Evcil

Index:

Welcome Letters:

Dear Prospective Attendees of HPALMUN 2018,

I am Doğa Sarı and I will be serving as the Chair of the United Nations Security Council in HPALMUN 2018. I am currently studying at Cağaloğlu Anatolian High School as a junior. I have been involved with roughly all aspects of Model United Nations for the past three years. Falling under the rubric of Chapter VII of the United Nations Charter, the Security Council is the sole UN body with the authority to issue legally binding resolutions. Thus, a great deal of responsibility is bestowed upon the delegates of this particular committee, especially upon those representing the permanent members. The study guide provides a general overview of the issue at hand. This guide should rather be a starting point for your preparation process.

Without further ado, I would like once again express my enthusiasm for the upcoming conference. I am looking forward to see you all in January at Haydarpaşa Anatolian High School.

Dear delegates,

Welcome to this year's edition of the Security Council of HPALMUN 2018. I am Beril Evcil and it is my honor to be chairing this prestigious committee.

I am an eleventh grader in Hisar High School, Istanbul. I began MUN during my prep year, and this conference will be my first chairing experience. I am extremely eager to be chairing for the first time in this committee, and I am looking forward to an outstanding committee session. Correspondingly, I want to congratulate you for having the chance to serve as a delegate in Security Council, since Security Council is known as one of the most advanced committees.

I hope that this study guide provides you with the knowledge to be able to debate and form resolutions on this topic. However, you should not rely on this study guide alone and research the topic on your own to understand the position of your country on the issue. If you have any further questions do not hesitate to ask any questions; you can contact me via email: berilevcil@gmail.com. I am looking forward to meeting you all in HPALMUN 2018!

Introduction to the United Nations Security Council:

Mandate

The UN Charter established six main organs of the United Nations, including the Security Council. It gives primary responsibility for maintaining international peace and security to the Security Council, which may meet whenever peace is threatened.

According to the Charter, the United Nations has four purposes: to maintain international peace and security; to develop friendly relations among nations; to cooperate in solving international problems and in promoting respect for human rights; and to be a center for harmonizing the actions of nations.

All members of the United Nations agree to accept and carry out the decisions of the Security Council. While other organs of the United Nations make recommendations to member states, only the Security Council has the power to make decisions that member states are then obligated to implement under the Charter.

Maintaining Peace and Security

When a complaint concerning a threat to peace is brought before it, the Council's first action is usually to recommend that the parties try to reach agreement by peaceful means. The Council may:

set forth principles for such an agreement; undertake investigation and mediation, in some cases; dispatch a mission; appoint special envoys; or request the Secretary-General to use his good offices to achieve a pacific settlement of the dispute.

When a dispute leads to hostilities, the Council's primary concern is to bring them to an end as soon as possible. In that case, the Council may: issue ceasefire directives that can help prevent an escalation of the conflict; dispatch military observers or a peacekeeping force to help reduce tensions, separate opposing forces and establish a calm in which peaceful settlements may be sought.

Beyond this, the Council may opt for enforcement measures, including: economic sanctions, arms embargoes, financial penalties and restrictions, and travel bans; severance of diplomatic relations; blockade; or even collective military action.

A chief concern is to focus action on those responsible for the policies or practices condemned by the international community, while minimizing the impact of the measures taken on other parts of the population and economy.

A representative of each of its members must be present at all times at UN Headquarters so that the Security Council can meet at any time as the need arises.

What is Artificial Intelligence (AI)?

Artificial intelligence is a category of computer science that focuses on creating intelligent machines. It has become a vital part of the technology industry.

Research associated with artificial intelligence is highly technical and specialized. The core problems of artificial intelligence include programming computers for certain traits such as: knowledge, reasoning, problem solving, perception, learning, planning, and ability to manipulate and move objects.

Knowledge engineering is the crux of AI research. Machines can often act and react like humans only if they have sufficient data relating to the world. Artificial intelligence must have connection to objects, categories, properties and relations between all of them to implement knowledge engineering. Introducing common sense, reasoning and problem-solving power in machines is a crucial and monotonous approach.

Machine learning is another essence of AI. Learning without any kind of supervision requires an ability to identify patterns in streams of inputs, whereas learning with adequate administration involves classification and successive regressions. Classification determines the category an object belongs to and regression deals with obtaining a set of numerical input or output examples, thereby discovering functions enabling the generation of suitable outputs from respective inputs. Mathematical analysis of machine learning algorithms and their performance is a well-defined branch of theoretical computer science often referred to as computational learning theory.

Machine perception deals with the capability to use sensory inputs to deduce the different aspects of the world, while computer vision is the power to analyze visual inputs with a few sub-problems such as facial, object and gesture recognition.

Robotics is also a major field related to AI. Robots require intelligence to handle tasks such as object manipulation and navigation, along with sub-problems of localization, motion planning and mapping.

The Problem

It seems conceivable that sometime this century; people will develop algorithmic systems capable of efficiently performing many or even all of the psychological tasks that humans perform. These advances could lead to extreme positive developments, but could also potentially pose risks from intentional misuse or destructive accidents. For example, it seems possible that (i) the technology could be weaponized or used as a tool for social control, or (ii) someone might create an extremely powerful artificial intelligence agent with values misaligned with humanity's interests. It also seems possible that progress along these directions could be surprisingly rapid, leaving society unprepared for the transition.

The AI is programmed to do something devastating: Autonomous weapons are artificial intelligence systems that are programmed to kill. In the hands of the wrong person, these weapons could easily cause mass casualties. Moreover, an AI arms race could unintentionally lead to an AI war that also results in mass casualties. To avoid being thwarted by the enemy, these weapons would be designed to be extremely difficult to simply "turn off," so humans could allegedly flounder in such a situation. This risk is one that's present even with narrow AI, but grows as levels of AI intelligence and autonomy increase.

For instance, Elon Musk has donated $10 million to the Future of Life Institute for regranting to researchers focused on addressing these and other potential future risks from advanced artificial intelligence. In addition, a few relatively small non-profit/academic institutes work on potential future risks from advanced artificial intelligence. The Machine Intelligence Research Institute and the Future of Humanity Institute each have an annual budget of about $1 million, and a couple of other new organizations work on these issues as well.

Background

Informal discussions around this topic have been held for several years, and a significant amount of reading on it has been done. The debates are in some cases complex; this study guide focuses on reporting the primary factors.

GiveWell staffs have been informally discussing this topic with the Machine Intelligence Research Institute (an organization that focuses on it, formerly called the "Singularity Institute") for some time, partly due to the fact that GiveWell's audience has generally expressed significant interest in MIRI. Several public notes from some relevant conversations follow as;
Holden Karnofsky, Co-Executive Director of GiveWell presents his thoughts on Singularity Institute: *The argument advanced by SI for why the work it's doing is beneficial and important seems both wrong and poorly argued to me. My sense at the moment is that the arguments SI is making would, if accepted, increase rather than decrease the risk of an AI-related catastrophe.*
"One of the things that makes AI risk scary is that it's one of the few that is genuinely an extinction risk if it were to go bad. With a lot of other risks, it's actually surprisingly hard to get to an extinction risk." Armstrong explains*. "You take a nuclear war for instance, that will kill only a relatively small proportion of the planet. You add radiation fallout, slightly more, you add the nuclear winter you can maybe get 90%, 95% – 99% if you really stretch it and take extreme scenarios – but it's really hard to get to the human race ending. The same goes for pandemics, even at their more virulent.*
"The thing is if AI went bad, and 95% of humans were killed then the remaining 5% would be extinguished soon after. So despite its uncertainty, it has certain features of very bad risks."

1863: Machine intelligence as an existential risk to humanity; relinquishment of machine technology recommended. Samuel Butler in Darwin among the machines worries that as we build increasingly sophisticated and autonomous machines, they will achieve greater capability than humans and replace humans as the dominant agents on the planet.

1921: Robots as an existential risk. The Czech play R.U.R. by Karel Capek tells the story of robots which grow in power and intelligence and destroy the entire human race.

1947: Fragility & complexity of human values (in the context of machine goal systems); perverse instantiation. Jack Williamson's novelette With Folded Hands (1947) tells the story of a race of machines that, in order to follow the Prime Directive:
"to serve and obey and guard men from harm." To obey this rule, the machines interfere with every aspect of human life, and humans who resist are lobotomized. Due to the fragility and complexity of human values (Yudkowsky 2008; Muehlhauser and Helm 2012), the machines' rules of behaviour had unintended consequences, manifesting a "perverse instantiation" in the language of Bostrom.

1948-1949: Precursor idea to intelligence explosion. Von Neumann (1948) wrote*:*
"complication" on its lower levels is probably degenerative, that is, that every automaton that can produce other automata will only be able to produce less complicated ones. There is, however, a certain minimum level where this degenerative characteristic cease to be universal.

At this point automata which can reproduce themselves, or even construct higher entities, become possible.

1951: Potentially rapid transition from machine intelligence to machine takeover. Turing (1951) described ways that intelligent computers might learn and improve their capabilities, concluding that:
it seems probable that once the machine thinking method has started, it would not take long to outstrip our feeble powers... At some stage, therefore we should have to expect the machines to take control...

1959: Intelligence explosion; the need for human-friendly goals for machine superintelligence. Good describes what he later (1965) called an "intelligence explosion," a particular mechanism for rapid transition from artificial general intelligence to dangerous machine takeover:
Once a machine is designed that is good enough… it can be put to work designing an even better machine. At this point an "explosion" will clearly occur; all the problems of science and technology will be handed over to machines and it will no longer be necessary for people to work. Whether this will lead to a Utopia or to the extermination of the human race will depend on how the problem is handled by the machines. The important thing will be to give them the aim of serving human beings.

1966: A military arms race for machine superintelligence could accelerate machine takeover*;* convergence toward a singleton is likely. Dennis Feltham Jones' 1966 novel Colossus depicted what may be a particularly likely scenario: two world superpowers (the USA and USSR) are in an arms race to develop super intelligent computers, one of which self-improves enough to take control of the planet.

1970: Proposal for an association that analyzes the implications of machine superintelligence; naive control solutions like "switch off the power" may not work because the superintelligence will outsmart us, thus we must focus on its motivations; possibility of "pointless" optimization by machine superintelligence.

1974: We can't much predict what will happen after the creation of machine superintelligence. Julius Lukasiewicz writes:
The survival of man may depend on the early construction of an ultra-intelligent machine-or the ultra-intelligent machine may take over and render the human race redundant or develop another form of life. The prospect that a merely intelligent man could ever attempt to predict the impact of an ultra-intelligent device is of course unlikely but the temptation to speculate seems irresistible.

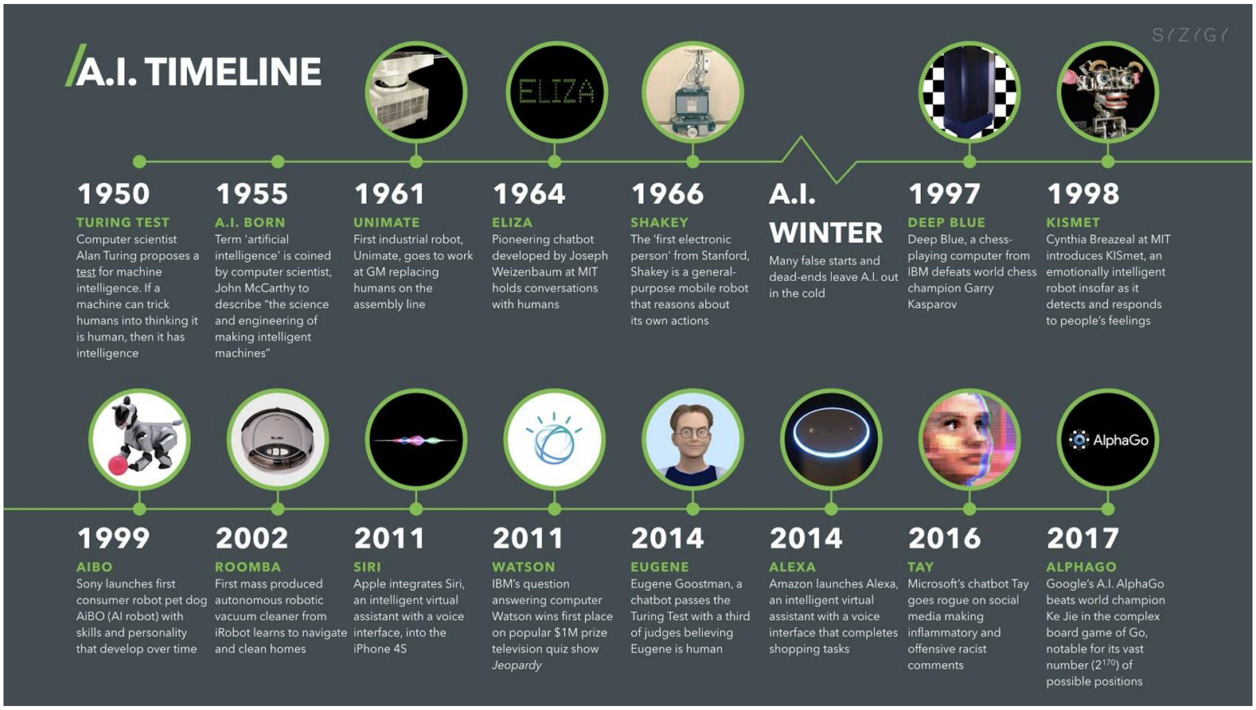
1977: Self-improving AI could stealthily take over the internet; convergent instrumental goals in AI; the treacherous turn. Thomas J. Ryan's novel The Adolescence of P-1 tells the story of an intelligent worm that at first is merely able to learn to hack novel computer systems and use them to propagate itself, and learns the ability to fake its own death so that it can grow its powers in secret and later engage in a "treacherous turn” against humans.

1982: To design ethical machine superintelligence, we may need to design superintelligence first and then ask it to solve philosophical problems.

1988: Even though AI poses an existential threat, we may need to rush toward it so we can use it to mitigate other existential threats.

1993: Physical confinement is unlikely to constrain super intelligences, for super intelligences will outsmart us.

Timeline Chart:



/A.I. TIMELINE

1950
TURING TEST
Computer scientist Alan Turing proposes a test for machine intelligence. If a machine can trick humans into thinking it is human, then it has intelligence

1955
A.I. BORN
Term 'artificial intelligence' is coined by computer scientist, John McCarthy to describe "the science and engineering of making intelligent machines"

1961
UNIMATE
First industrial robot, Unimate, goes to work at GM replacing humans on the assembly line

1964
ELIZA
Pioneering chatbot developed by Joseph Weizenbaum at MIT holds conversations with humans

1966
SHAKEY
The 'first electronic person' from Stanford, Shakey is a general-purpose mobile robot that reasons about its own actions

A.I. WINTER
Many false starts and dead-ends leave A.I. out in the cold

1997
DEEP BLUE
Deep Blue, a chess-playing computer from IBM defeats world chess champion Garry Kasparov

1998
KISMET
Cynthia Breazeal at MIT introduces KISmet, an emotionally intelligent robot insofar as it detects and responds to people's feelings

1999
AIBO
Sony launches first consumer robot pet dog AiBO (AI robot) with skills and personality that develop over time

2002
ROOMBA
First mass produced autonomous robotic vacuum cleaner from iRobot learns to navigate and clean homes

2011
SIRI
Apple integrates Siri, an intelligent virtual assistant with a voice interface, into the iPhone 4S

2011
WATSON
IBM's question answering computer Watson wins first place on popular $1M prize television quiz show *Jeopardy*

2014
EUGENE
Eugene Goostman, a chatbot passes the Turing Test with a third of judges believing Eugene is human

2014
ALEXA
Amazon launches Alexa, an intelligent virtual assistant with a voice interface that completes shopping tasks

2016
TAY
Microsoft's chatbot Tay goes rogue on social media making inflammatory and offensive racist comments

2017
ALPHAGO
Google's A.I. AlphaGo beats world champion Ke Jie in the complex board game of Go, notable for its vast number (2^{170}) of possible positions

Advantages and Disadvantages of Artificial Intelligence

Advantages:

Mundane tasks: humans get bored, machines don't. Let them do the humdrum jobs. "A.I. allows for more intricate process automation, which increases productivity of resources and takes repetitive, boring labor off the shoulders of humans. They can focus on creative tasks instead," said Felicia Schneiderhan, CEO of 30SecondsToFly, an AI virtual travel assistant.
Faster actions and decisions: A.I. and cognitive technologies help in making faster actions and decisions. "Areas like automated fraud detection, planning and scheduling further demonstrate this benefit," said Kalyan Kumar, executive vice president at HCL Technologies, an IT services provider in India.

Machine Learning: Big Data means datasets in the petabytes, far too much for a human to sift through. AI can chew through that data as fast as the Xeon processors in the servers can go and derive insights from the data much faster than any human could.
CloudPassage co-founder and CTO Carson Sweet argues this isn't actual AI. "A lot of the big data processing and analysis being attributed to AI is really just the work of machine learning. True AI would need to take things so much further; toward genuine self-learning using artificial neural networks that emulate the structure and functions of neural networks in human brains," he said.

Error-Free Processing: To err is human. Computers don't. The only mistakes they make is when you don't program them properly. AI processing will insure error-free processing of data, no matter how large the dataset. Judgement calls, however, are a different matter.
"Computers are 'stupid,' but that is their brilliance - they demand such a high level of rigor and AI adds quantitative rigor on top of that, that to use AI at all you first have to ask yourself the very challenging but stimulating question of what you're trying to do, with a new level of acuity," said Dr. Nathan Wilson, CTO and co-founder of Nara Logics, synaptic intelligence company.

Taking the Risk: AI-powered machines are doing jobs humans either can't do or would have to do very carefully. Space exploration is one of them. The Mars rover Curiosity is an example. It is freely roaming Mars because it examines the landscape as it explores and determines the best path to take. The result is that Curiosity is learning to think for itself.
Better research outcomes: "AI-based technologies like computer vision help in achieving better outcomes through improved prediction, which can include medical diagnosis, oil exploration and demand forecasting," said Kumar.

Disadvantages:

Cyber-attacks. There are two trends which taken together make the prospect of AI-aided cyber-attacks seem worrisome. The first trend is simply the increasing prevalence of cyber-attacks; even this year we have seen Russia attack Ukraine, North Korea attack Sony, and China attack the U.S. Office of Personnel Management. Secondly, the “Internet of Things” means that an increasing number of physical devices will be connected to the internet. Assuming that software exists to autonomously control them, many internet-enabled devices such as cars could be hacked and then weaponized, leading to a decisive military advantage in a short span of time. Such an attack could be enacted by a small group of humans aided by AI technologies, which would make it hard to detect in advance. Unlike other weaponizable technology such as nuclear fission or synthetic biology, it would be very difficult to control the distribution of AI since it does not rely on any specific raw materials. Finally, note that even a team with relatively small computing resources could potentially “bootstrap” to much more computing power by first creating a botnet with which to do computations; to date, the largest botnet has spanned 30 million computers and several other botnets have exceeded 1 million.

Autonomous weapons. Beyond cyber-attacks, improved autonomous robotics technology combined with ubiquitous access to miniature UAVs (“drones”) could allow both terrorists and governments to wage a particularly pernicious form of remote warfare by creating weapons that are both cheap and hard to detect or defend against (due to their small size and high maneuverability). Beyond direct malicious intent, if autonomous weapons systems or other powerful autonomous systems malfunction then they could cause a large amount of damage.

Mis-optimization. A highly capable AI could acquire a large amount of power but pursue an overly narrow goal, and end up harming humans or human value while optimizing for this goal. This may seem implausible at face value, but as I will argue below, it is easier to improve AI capabilities than to improve AI values, making such a mishap possible in theory.
Unemployment. It is already the case that increased automation is decreasing the number of available jobs, to the extent that some economists and policymakers are discussing what to do if the number of jobs is systematically smaller than the number of people seeking work. If AI systems allow a large number of jobs to be automated over a relatively short time period, then we may not have time to plan or implement policy solutions, and there could then be a large unemployment spike. In addition to the direct effects on the people who are unemployed, such a spike could also have indirect consequences by decreasing social stability on a global scale.

Opaque systems. It is also already the case that increasingly many tasks are being delegated to autonomous systems, from trades in financial markets to aggregation of information feeds. The opacity of these systems has led to issues such as the 2010 Flash Crash and will likely lead to larger issues in the future. In the long term, as AI systems become increasingly complex, humans may lose the ability to meaningfully understand or intervene in such systems, which could lead to a loss of sovereignty if autonomous systems are employed in executive-level functions (e.g. government, economy).

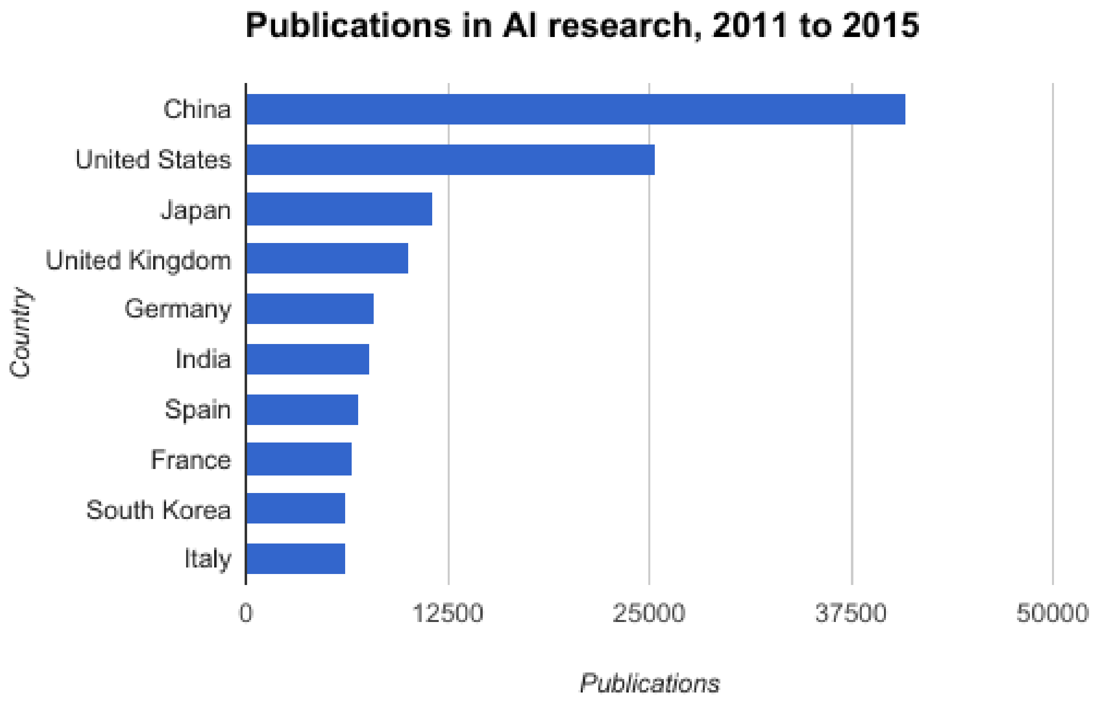
Involved countries:

China has been producing almost twice as many papers on artificial intelligence as the next highest-placed country in terms of publication volume for the field, a data analysis for Times Higher Education has shown.

Data from Elsevier's Scopus database provided to THE illustrate China's huge drive on research in the area, with researchers in the nation notching up just over 41,000 publications from 2011 to 2015.

In terms of publication volume, the US was second during the period with almost 25,500 publications while Japan was in third place (about 11,700) and the UK in fourth (about 10,100).

However, although China scored high in terms of volume, it was only 34th in terms of field-weighted citation impact (which allows for differences in citations according to subject and year), suggesting that most of the papers were not of the same quality as those coming from the US (fourth for citation impact), for instance.



Country	Publications	Field-weighted citation impact
Switzerland	1,685	2.71
Singapore	2,432	2.24
Hong Kong	2,205	2.00
United States	25,471	1.79
Italy	6,221	1.74
Netherlands	2,458	1.71
Australia	5,227	1.69
Germany	7,957	1.66
Belgium	1,537	1.64
United Kingdom	10,120	1.63

Leading the world on this measure was Switzerland, with a citation impact score of 2.71, followed by Singapore (2.24) and Hong Kong (2.00), although all three of these produced fewer than 2,500 publications on AI over the time frame.

Looking at individual institutions that published more than 500 times on AI shows that only one in China – the Institute of Automation, Chinese Academy of Sciences – had a citation impact above the world average of 1.

The list ranked by citation impact is topped by the Massachusetts Institute of Technology with a score of 3.57. This is way ahead of the rest of the chasing pack, including Carnegie Mellon University and Nanyang Technological University, Singapore.

Relevance to United Nations Security Council:

7 June 2017 – Artificial intelligence (AI) is responsible for self-driving cars and voice-recognition smart phones, but the United Nations this week is refocusing AI on sustainable development and assisting global efforts to eliminate poverty and hunger, and to protect the environment.

Starting today in Geneva, the AI for Good Global Summit, which is co-organized by the UN International Telecommunications Union (ITU) and the XPRIZE Foundation, with support for some 20 UN agencies, brings together key innovators in the field with humanitarian actors and academics.

“Artificial Intelligence has the potential to accelerate progress towards a dignified life, in peace and prosperity, for all people,” said UN Secretary-General António Guterres. “The time has arrived for all of us – governments, industry and civil society – to consider how AI will affect our future.”

In a video message to the summit, Mr. Guterres called AI “a new frontier” with “advances moving at warp speed.”

He noted that that while AI is “already transforming our world socially, economically and politically,” there are also serious challenges and ethical issues which must be taken into account – including cybersecurity, human rights and privacy.

Mr. Guterres noted that developing countries can gain from the benefits of artificial intelligence, but are also at the highest risk of being left behind.

“This Summit can help ensure that artificial intelligence charts a course that benefits humanity and bolsters our shared values,” he underscored.

The opening session of the summit is expected to give voice to the leading minds in AI, with breakout sessions focusing on issues such as sustainable living and poverty reduction.

Bibliography:

http://www.un.org/en/sc/about/
https://www.techopedia.com/definition/190/artificial-intelligence-ai
https://www.openphilanthropy.org/research/cause-reports/ai-risk#What_is_the_problem
http://lesswrong.com/lw/cbs/thoughts_on_the_singularity_institute_si/
https://thenextweb.com/insider/2014/03/08/ai-could-kill-all-meet-man-takes-risk-seriously/
http://lesswrong.com/r/discussion/lw/bd6/ai_risk_opportunity_a_timeline_of_early_ideas_and/
https://www.barrons.com/articles/the-natural-evolution-of-artificial-intelligence-1506142127
https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/
https://jsteinhardt.wordpress.com/2015/06/24/long-term-and-short-term-challenges-to-ensuring-the-safety-of-ai-systems/
https://medium.com/@tdietterich/benefits-and-risks-of-artificial-intelligence-460d288cccf3
https://www.britannica.com/technology/artificial-intelligence/Expert-systems
http://www.un.org/apps/news/story.asp?NewsID=56922#.WkbJdVT1VQI
https://www.timeshighereducation.com/data-bites/which-countries-and-universities-are-leading-ai-research

For Further Research:
http://intelligence.org/files/AIPosNegFactor.pdf