the 2022
innovations dialogue.

AI DISRUPTION, PEACE & SECURITY

Conference Report

Prepared by Alisha Anand and Wenting He

UNIDIR UNITED NATIONS INSTITUTE FOR DISARMAMENT RESEARCH

ACKNOWLEDGEMENTS

Support from UNIDIR's core funders provides the foundation for all the Institute's activities. The 2022 Innovations Dialogue was organized by the Security and Technology Programme, which is funded by the governments of Czechia, Germany, Italy, the Netherlands and Switzerland, and by Microsoft.

The 2022 Innovations Dialogue was the fourth edition of one of UNIDIR's flagship events. UNIDIR would like to thank all the speakers, moderators and participants for their presentations, comments and contributions at the 2022 Innovations Dialogue, which are the basis of this report. Details about speakers and moderators are given in the conference agenda enclosed in the report.

ABOUT UNIDIR

The United Nations Institute for Disarmament Research (UNIDIR) is a voluntarily funded, autonomous institute within the United Nations. One of the few policy institutes worldwide focusing on disarmament, UNIDIR generates knowledge and promotes dialogue and action on disarmament and security. Based in Geneva, UNIDIR assists the international community to develop the practical, innovative ideas needed to find solutions to critical security problems.

CITATION

A. Anand and W. He, *The 2022 Innovations Dialogue: AI Disruption, Peace and Security,* Conference Report, Geneva, Switzerland: UNIDIR, 2023.

NOTE

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries. The views expressed in the publication are the sole responsibility of the individual authors. They do not necessary reflect the views or opinions of the United Nations, UNIDIR, its staff members or sponsors.

ABOUT THE SECURITY AND TECHNOLOGY PROGRAMME

Contemporary developments in science and technology present new opportunities as well as challenges to international security and disarmament. UNIDIR's Security and Technology Programme (SecTec) seeks to build knowledge and awareness on the international security implications and risks of specific technological innovations and convenes stakeholders to explore ideas and develop new thinking on ways to address them.

ABOUT THE AUTHORS

Alisha Anand is an Associate Researcher in the Security and Technology Programme at UNIDIR. Her work is focused on the international security implications of new and emerging technologies and on technology governance, particularly in the field of artificial intelligence. Before joining UNIDIR, she worked on non-proliferation and export controls with the Disarmament and International Security Affairs Division of the Indian Ministry of External Affairs, the Manohar Parrikar Institute for Defence Studies and Analyses, and the Federation of Indian Chambers of Commerce & Industry. She holds a master's degree in law and diplomacy from the Fletcher School, Tufts University, where she specialized in international security and international law. Follow Alisha on Twitter @AlishaAnand912

Wenting He is the Programme and Events Assistant in UNIDIR's Security and Technology Programme, where she provides administrative, operational and event support for all workstreams. Before joining UNIDIR, she worked for the United Nations Office at Geneva, assisting with programme implementation, and was a research assistant with the Global Initiative Against Transnational Organized Crime. She holds a master's degree in international affairs from the Graduate Institute of International and Development Studies, Geneva, and a bachelor's degree in diplomacy from China Foreign Affairs University, Beijing.

CONTENTS

ABOUT THE SECURITY AND TECHNOLOGY PROGRAMME 3

ABOUT THE AUTHORS 3

ABBREVIATIONS AND ACRONYMS 5

ABOUT THE INNOVATIONS DIALOGUE 6

THE 2022 INNOVATIONS DIALOGUE: AI DISRUPTION, PEACE AND SECURITY 7

HIGHLIGHTS 11

PART I: AI AND ITS STATE OF PLAY 14

PART II: THE DISRUPTIVE IMPACT OF AI ON INTERNATIONAL PEACE AND SECURITY 18

- Uses of AI in Military Operations 18
- Disruptive Impact of AI Across Domains of Warfare 22
- AI for Peace – AI and Conflict Prevention and Peacebuilding 26

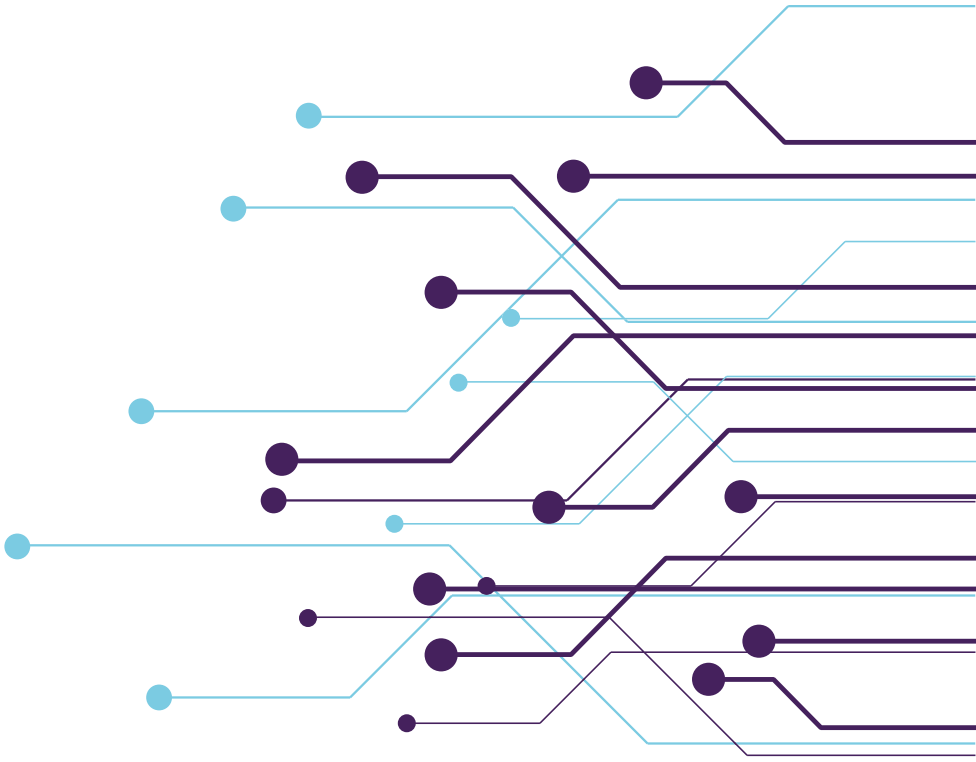
PART III: TOWARDS RESPONSIBLE AI 30

REFERENCE LIST 37

CONFERENCE AGENDA 42

ABBREVIATIONS AND ACRONYMS

AI	Artificial intelligence
DoD	Department of Defense (United States)
DPPA	Department of Political and Peacebuilding Affairs (United Nations)
EIA	Ethical impact assessment
ISR	Intelligence, surveillance and reconnaissance
OODA	Observe–orient–decide–act
RAI	Responsible AI
UNESCO	United Nations Educational, Scientific and Cultural Organization



ABOUT THE INNOVATIONS DIALOGUE

Launched in 2019, the Innovations Dialogue is one of UNIDIR's flagship events. The conference series was established pursuant to the 2018 General Assembly resolution on the "Role of science and technology in the context of international security and disarmament".[1] The Innovations Dialogue provides a unique multi-stakeholder forum – convening experts from the diplomatic and policy community, technical and scientific community, industry groups, and academia and civil society – to collectively examine developments in science and technology that have potentially radical and novel implications for international peace and security and for disarmament. Through fact-based and balanced discussions, the dialogue aims to dispel myths about scientific and technological innovations and build a shared understanding of the potential benefits, risks and policy challenges posed by such innovations.

The Secretary-General's May 2018 Agenda for Disarmament, "Securing Our Common Future", and his 2018 and 2021 reports on "Current developments in science and technology and their potential impact on international security and disarmament efforts" recognize UNIDIR's role as a source of knowledge and ideas, as well as a convener of multi-stakeholder dialogues, at the nexus of technology and security.[2]

The key objectives of the Innovations Dialogue are:

- **To collaboratively examine beneficial applications** of advances in science and technology for international peace and security **as well as new and converging challenges or risks** that arise.
- **To promote multi-stakeholder engagement and build new relationships** among a range of actors and tools that can contribute to mitigating potential harms, harnessing potential benefits and promoting responsible innovation.
- **To explore how multi-stakeholder dialogue can facilitate policy responses** to developments in science and technology that have potentially radical and novel implications for international security and disarmament, with a view to identifying gaps or opportunities where early thinking on strategies for risk mitigation may be beneficial.

1 UNGA (2018b).
2 United Nations Office for Disarmament Affairs (2018); UNGA (2018a); UNGA (2021).

THE 2022 INNOVATIONS DIALOGUE: AI DISRUPTION, PEACE AND SECURITY

Stage preparations before the 2022 Innovations Dialogue

Continuous and novel advances in artificial intelligence (AI) and efforts to integrate AI technologies in critical sectors are gradually transforming all aspects of our society. However, as the field of AI is evolving rapidly, there is conceptual ambiguity and uncertainty regarding what AI is, what it can do, its perils and promises, and where it is headed. This makes the governance of AI technologies challenging, particularly in the high-risk context of defence. These challenges are further compounded by the nature of the AI research and innovation landscape. It has a strong open-source and democratized culture and is driven largely by the AI research community comprising of big technology companies, start-ups, university laboratories and individual AI researchers.

Thus, states alone cannot grasp and address the complex issues associated with advances in AI technologies and their impact on international peace and security. These complex challenges first and foremost require systematic and continuous multi-stakeholder deliberations. In this spirit, **the 2022 Innovations Dialogue convened representatives from governments, the AI research community and civil society to collectively decode AI and examine the disruptive impact of AI advances on international peace and security today and in the future.**

The Dialogue brought together 26 expert speakers from governments, international organizations, academia and industry and nearly 1,700 (virtual and in-person) participants from around the world.[3] Together, they decoded the concept of AI and the state of play of AI technologies, while reflecting on the current obstacles to and opportunities for advancement of AI in the future. The Dialogue also examined the potential risks and benefits of the use of AI in the context of international peace and security, including in military operations across domains of warfare, and for conflict prevention and peacebuilding. Finally, the Dialogue unpacked what Responsible AI (RAI) is, how it relates to international peace and security, and how it can be put into practice. In doing so, it also reflected on the shared roles and responsibilities of key stakeholders with respect to building an RAI culture.

This report provides a summary of the key themes, issues and takeaways that emerged from the 2022 Innovations Dialogue. Based on the discussions that took place, Part I of the report seeks to provide a foundational understanding of the concept of AI and its state of play. Part II examines the disruptive impact of AI on international peace and security. In particular, it discusses the risks and benefits of uses of AI in military operations and across domains of warfare as well as the opportunities and challenges of harnessing AI technologies for conflict prevention and peacebuilding. Part III of the report examines the path to Responsible AI. It unpacks the RAI governance approach and discusses how it is and can be operationalized. It also reflects on the value of building an RAI culture.

[3] The 26 speakers – 14 female and 12 male – represented 15 nationalities.

UNIDIR Director Robin Geiss opening the 2022 Innovations Dialogue

A participant taking a picture of the conference agenda

Panel on What even is, AI? – The State of Play and the Future of AI, moderated by Ioana Puscas, UNIDIR

Giacomo Persi Paoli, UNIDIR

Panel on Uses of AI in Military Operations, moderated by Alisha Anand, UNIDIR

Participants taking note during a session

Discussions during a coffee break

Panel on the Disruptive Impact of AI Across Domains of Warfare, moderated by Beyza Unal, UNODA

Panel on AI for Peace – AI and Conflict Prevention and Peacebuilding, moderated by Sarah Grand-Clément, UNIDIR

Attendees at the 2022 Innovations Dialogue

Speakers on the stage

Panel on Unpacking and Operationalizing the Responsible AI Toolbox, moderated by Giacomo Persi Paoli, UNIDIR

Concluding panel on Building a Culture of Responsible AI: A Shared Responsibility, moderated by Robin Geiss, UNIDIR

UNIDIR team hosting the 2022 Innovations Dialogue in New York

The video recording of the conference is available on UNIDIR's website here.

HIGHLIGHTS

AI and its state of play

- At present there is no widely accepted definition for AI as **AI is a broad discipline that is defined in different ways for different purposes**.
- AI can simply be thought of **as a system in which algorithms use data to make decisions** (or perform tasks) on our behalf or help humans make decisions (or perform tasks).
- **AI systems are what we make them**, and they will be what we want them to be. Essentially, they are human artifacts or human constructs designed by humans and trained largely on socially generated data.
- As the field of AI is evolving rapidly, it is becoming apparent that **AI presents unprecedented opportunities** to augment human capabilities, particularly in problem-solving and decision-making. However, at the same time, **significant ethical, legal, safety and security concerns** remain and are coming to the fore as AI systems are increasingly adopted across sectors, including the defence sector.
- **These concerns range across issues** related, but not limited, to transparency, reliability, predictability, understandability, accountability, bias and discrimination, and technical robustness.
- **A key issue that causes AI systems to make errors or fail is that AI can be biased.** There are three main dimensions of bias – *pre-existing bias*, which concerns bias in data, *technical bias*, which is introduced by the operation of the technical system itself and may amplify pre-existing bias, and *emergent bias*, which arises in the context of use of a system.
- **Deciding how much autonomy should be given to an AI system to perform which tasks in which contexts is crucial** because errors or failures in performing safety- and security-critical tasks can have adverse consequences for individuals, organizations and societies.
- **At present, AI systems are becoming good at performing narrow and specific, well-defined tasks** that are often repeatable and have clear criteria for success, on the basis of which developers and users can judge whether the system has achieved its purpose.
- **AI systems nonetheless remain brittle when it comes to their performance in dynamic and cluttered environments**, where these systems encounter uncertain conditions.
- While it is hard to predict the exact trajectory of AI advancements, what is becoming evident is that the **future will witness new forms and structures of collaboration and coordination between humans and AI systems**.

HIGHLIGHTS

The disruptive impact of AI on international peace and security

- The steady integration of AI technologies into an increasing number of military applications could **transform the conduct of military operations by enhancing military capabilities** in terms of efficiency, speed, precision, survivability and coordination.
- **In the military domain, at present the uses of AI technologies are rudimentary and not at scale**, but they are perhaps groundbreaking in the sense that they have not been attempted before. The most prominent uses of AI technologies today are in intelligence, surveillance and reconnaissance (ISR) operations, strengthening cyber defences, conducting as well as mitigating AI-enhanced influence operations, and enhancing combat simulations for military training and planning.
- While AI technologies could offer benefits in military operations, they also create unique risks. For example, **increased speed in military decision-making could result in miscalculation and inadvertent escalation**. Current AI technologies are also brittle and prone to making errors and being fooled by adversarial spoofing or hacking. Due to this, they may ultimately be less accurate and precise than human operators in complex battlefields.
- As governments around the world are increasingly seeking to harness AI technologies across sectors, including in the military domain, they are developing national AI strategies and even defence-specific AI strategies. To ensure that such high-level strategy documents can have the desired operational impact, governments must remain cognizant of three key considerations – **AI is all about trade-offs, AI innovation involves uncertainty, and not every nail needs an AI-enabled hammer**.
- Given that the civilian AI industries work for or with militaries to build AI systems, it is imperative that governments and their militaries, regardless of the AI governance approach they adopt, take measures **to ensure that civilian and military governance frameworks align with respect to military applications of AI technologies**.
- As an enabling technology, the integration of **AI technology across domains of warfare – from cyber and biological to nuclear, and especially in convergence with other powerful dual-use technologies** – can have benefits for international peace and security as well as pose novel risks.
- **AI advancements can be harnessed to build and sustain peace**. In recent years, United Nations agencies have explored and even deployed AI-enabled applications for conflict prevention and peacebuilding around the globe, including to facilitate dialogue among different communities and better understand the different needs and concerns within a local context. However, there remain practical implementation challenges to deploying AI solutions for conflict prevention and peacebuilding at scale.

HIGHLIGHTS

Towards Responsible AI

- Given the ethical, legal, safety and security concerns that AI technologies present, **governments, intergovernmental organizations, private sector entities and members of civil society are developing normative instruments such as principles and standards to guide the AI system lifecycle**. These aim to ensure that AI systems are researched, designed, developed, deployed and used in a responsible manner in accordance with legal requirements and ethical values. This approach to AI governance is broadly known as Responsible AI.
- **RAI can be understood as a principles-based, socio-technical approach to the research, design, development, deployment, use, maintenance and governance of AI systems** across sectors that is conscious of and considers the effects (both positive and negative) that such systems may have on individuals, communities and society at large.
- **This RAI approach** helps to prevent unethical or irresponsible applications of AI technologies and consequently to **build trust in AI systems. The trust in turn is an enabler for the rapid adoption and deployment of AI systems**.
- Through ethical principles or guidelines and a combination of tools such as (but not limited to) testing standards, risk-assessment frameworks, conformity-assessment schemes, accountability checks and employment guidance, the **RAI approach can proactively ensure that decisions made through the AI system lifecycle result in intended outcomes**.
- Since Responsible AI is a lifecycle approach towards managing risks and preventing possible harms while facilitating responsible use, **stakeholders involved and concerned with every stage of the AI system lifecycle, from research to use, have a shared role to play**.
- **RAI efforts usually begin with the adoption of broad AI principles or guidelines** that encompass technical, legal and ethical requirements that AI systems should meet in order to be responsible and trustworthy. Committing to principles is, however, not sufficient to achieve responsible and trustworthy AI.
- **Broad principles need to translate into practice.** Thus, beyond the commitment to principles, governments and organizations that create or use AI, at their own levels, should develop detailed practical guidance for AI actors involved in the AI system life-cycle and put in place tools, processes, and governance structures and mechanisms for the operationalization of AI principles.
- **Scaling up RAI practices** and realizing the adoption of responsible and trusted AI systems ultimately **requires cultivating and sustaining a culture of Responsible AI** in which RAI-related considerations and values are instilled in the organizational culture and viewed as an integral and enabling part of AI development, rather than barriers to it, at both the system-wide and individual levels.

PART I: AI AND ITS STATE OF PLAY[4]

Continuous advances in the field of AI and machine learning and efforts to integrate AI technologies in critical sectors are gradually transforming all aspects of our society, and the defence and security sector is no exception. But, **as the field of AI is evolving rapidly, there is conceptual ambiguity and uncertainty regarding what AI is, what it can do, its perils and promises, and where it is headed**. This session sought to provide a foundational understanding of the concept of AI and its state of play.

Presently there is no widely accepted definition for AI, perhaps because AI is a very broad discipline that is defined in different ways for different purposes since what AI is used for depends on the domain of use.[5] Broadly, AI could be described as the science and engineering of making "intelligent" machines, where intelligence entails the ability to perform well at goal-oriented tasks and to exhibit behaviours to achieve those by interacting within a dynamic environment in which there are uncertain conditions and only partially visible information.[6]

AI can also be simply thought of as a system in which algorithms use data to make decisions (or perform tasks) on our behalf or help humans make decisions (or perform tasks). A simple analogy of the process of baking bread can be used to understand the three main components – algorithms, data and decisions.[7] An algorithm is simply a sequence of steps – it entails the steps that need to be taken to transform the ingredients into a loaf of bread. The algorithm can be fully prescribed in the sense that it may list exactly what ingredients are needed, what quantity of each ingredient is needed, in what order the ingredients should be combined, at what temperature the bread should be baked, and so on. Such algorithms are rules based. If the rules are pre-determined and well-defined, then the algorithm can always be executed to get roughly the same output. Alternatively, algorithms can also be adaptive. Using the same analogy of baking bread, a "learning" algorithm can learn the recipe to bake bread from our experience of what a good loaf of bread tastes like. The structure of the recipe may remain the same, but we may try different combinations of ingredients, temperatures and baking times and then assess the taste of the different bread loafs. Based on those assessments, we can determine what combination of parameters work well together. In other words, by leveraging data to improve performance, such algorithms can learn and adapt to give an output without following explicit instructions. This subfield of AI is known as machine learning.

The data used to train algorithms comes in multiple forms. They include input data, which is the ingredients needed in the recipe. Another form of data is parameters, such as oven temperatures and baking times. The third form is data that describes the output, which are objectively measurable factors – for example, the weight of the bread loaf or its nutritional

---

4 This part summarizes the discussions that took place during the segment "Panel I: What even is AI? – The state of play and the future of AI"; https://www.unidir.org/ID22. This panel was comprised of Abhishek Gupta (Founder and Principal Researcher, Montreal AI Ethics Institute and Senior Responsible AI Leader & Expert, Boston Consulting Group), Jason Lin (Research Fellow in AI Safety, Stanford Existential Risk Initiative; 3D Perception Lead, Lyft Self-Driving; Autonomous AI, Google X) and Julia Stoyanovich (Associate Professor of Computer Science & Engineering, Associate Professor of Data Science, Director of the Center for Responsible AI, New York University).

5 Stoyanovich (2022).

6 Gupta (2022).

7 The next few paras largely summarize Julia Stoyanovich's presentation at the panel discussion. Stoyanovich (2022).

value. The fourth kind of data is human judgement or the subjective feedback that humans give on the output, which often is more important in the "learning" process than objective properties of the output.

With regard to decisions, humans need to make critical decisions after every execution of the algorithm to ensure that it is able to generate the desired output. In the bread-baking analogy, humans make important decisions after every step. They must consider: Is the loaf tasty? Among the different recipes, which should be considered a success? Should certain specific ingredients or combination of ingredients or parameters always or never be used? There are some even more consequential decisions: Have enough recipes been tried to pass on the experience to an AI system to bake bread on our behalf? Can the machine also be trusted to bake a different type of bread based on "learning" to bake one type? And most importantly, should we let machines make judgements on behalf of humans, deciding which baked goods came out well and which did not?

Deciding how much autonomy should be given to a learning algorithm and to perform which tasks in which contexts is crucial because errors or failures in performing safety- and security-critical tasks can have adverse consequences for individuals, organizations and societies. While errors occurring in AI-based spam filters or AI-enabled video and computer games may have low impact, there are recent real-world instances which show that mistakes can result in grave, irreversible harms, and even the loss of human life – from a self-driving car killing a pedestrian to perpetuated or developing new biases in AI-based recruitment tools.[8] Asking AI systems to perform tasks that are complex and difficult (even for humans in some cases), such as predicting whether someone will perform well in a job if they are hired based on past data, entails a reasonable level of risk that the system will make mistakes. Similarly, in certain high-risk military contexts, AI system errors or failures could have catastrophic consequences, resulting in the loss of civilian life or damage to critical infrastructure.

A key issue that causes AI systems to make errors or fail is that AI can be biased. **There are three main dimensions of bias – *pre-existing bias*, which concerns bias in data, *technical bias*, which is introduced by the operation of the technical system itself and may amplify *pre-existing bias*, and emergent bias, which arises in the context of use of a system.**[9] First, with respect to pre-existing bias, we attempt to mirror the world in the data sets that we build. However, that reflection of the world can be distorted – we may under-, over- or mis-represent particular parts or facets of the world in the data on which an algorithm is trained. For instance, an autonomous car that fails to recognize the presence of a pedestrian in a wheelchair (an object it is expected to encounter) could do so because its object-recognition algorithm was trained on a data set that did not account for pedestrians in wheelchairs. Similar concerns are raised about the ability of weapon systems with autonomous functions to make nuanced discrimination between combatants and non-combatants, such as whether a combatant incapacitated by injuries is a lawful target.

8 McKendrick and Thurai (2022).
9 Friedman and Nissenbaum (1996).

A second cause of pre-existing bias is when training data sets that would be representative of the intended context of use are simply unavailable. For example, in the case of an autonomous weapon, this could be a specific type of battlefield with a specific kind of terrain, weather conditions or other environmental factors. Beyond the issue of data representation, pre-existing bias can occur because, even if the world can be perfectly mirrored in the data, it would still be a reflection of the world we live in and not what it could or should be – biases that pre-exist in our world such as gender or racial bias can be perpetuated in the data on which algorithms are trained. This could have grave consequences from an ethical or legal perspective. Even the most state-of-the-art AI models are prone to such biases. This includes novel text-to-image generation models like DALL-E 2 developed by OpenAI – early tests of the model as part of OpenAI's red teaming process showed that the model leaned towards generating images of white men, overly sexualizing women and reinforcing racial stereotypes in the images it generated.[10]

Second, with respect to technical bias, the properties of the technical system may itself exacerbate inequities that exist in the world. Such biases usually emerge during the design phase of the algorithm. This could, for instance, occur as a result of how the objectives for the system are stated. For example, in the case of an autonomous weapon system that is tasked to classify individuals as civilians (the negative class) or combatants (the positive class), the criteria we use to determine whether the system can perform this task can introduce bias. Would we consider the system to work well if it never misclassifies a civilian (i.e., having a low false positive rate) or if it never misclassifies a combatant (i.e., having a low false negative rate)? Unless the weapon system has perfect accuracy, it may not be able to simultaneously achieve perfect performance according to both these goals. Therefore, it is crucial to determine which goals to prioritize and why. Technical bias can also emerge when system designers attempt to make human aspects machine readable or, in other words, quantify what is fundamentally qualitative, such as human emotions.[11]

Third, with respect to emergent bias, biases may emerge due to the interaction between the AI system and its users. Such biases do not pre-exist in the training data but arise in the context of the use of a system. A prominent real-world example of this is the AI chatbot Tay, which was developed by Microsoft. The bot was an experiment in conversational AI, designed to mimic a teenage girl by learning their style of communicating and slang through interaction with human users on Twitter. Within less than a day of interactions with human users on Twitter who tweeted racist comments, the chatbot mimicked the human users and started tweeting racist content.[12]

10 K. Johnson (2022).
11 Von Laufenberg (2020).
12 Schwartz (2019).

Other forms of bias relating to human–machine interaction that can result in errors or failure are automation bias and algorithmic aversion. Automation bias is when humans become over-reliant on AI systems. They get fatigued into always trusting the AI system's outputs or predictions despite the availability of contradictory information, even if the latter is correct. Algorithmic aversion, on the other hand, is not trusting AI systems enough or at all. This could cause users to ignore valuable inputs provided by AI systems that could improve decision-making.[13]

Can we mitigate AI system errors and failures? **AI systems are what we make them, and they will be what we want them to be.** Essentially, they are human artifacts or human constructs designed by humans and trained largely on socially generated data. Rigorous testing and evaluation are essential throughout the research, design, development, deployment and use phases of AI systems to ensure that they do not replicate human biases more profoundly and at scale, perform reliably and safely in normal as well as unanticipated circumstances, and evolve in a manner that is consistent with original expectations. Moreover, developers and users of AI systems need to carefully evaluate whether, when and how to delegate decisions and actions to AI systems. Ultimately, **any decision or prediction made by an AI system should always be accompanied by human judgment and oversight, and users should be equipped to override the decisions or predictions where necessary**.

At present, AI systems are becoming good at performing narrow and specific, well-defined tasks that are often repeatable and have clear criteria for success on the basis of which developers and users can judge whether the system has achieved its purpose. A recent prominent example of this is DeepMind's AlphaFold, which is a state-of-the-art AI system that can generate predictions of protein structures with unprecedented accuracy and speed. This addresses one of the fundamental challenges in biology – to understand the building blocks of cells and enable quicker and more advanced drug discovery.[14] **However, AI systems remain brittle when performing in dynamic and cluttered environments, where these systems encounter uncertain conditions**.[15] A notable example of this is autonomous driving. It has been challenging to develop and deploy fully self-driving vehicles because such systems have to not only operate in complex, dynamic environments comprised of multiple roadways, street signs, pedestrians, other vehicles, buildings and so on, but also predict inherently unpredictable human behaviour such as the behaviour of any pedestrians they may encounter.[16]

Nevertheless, the field of AI is advancing at a rapid pace. It is possibly moving towards progress with respect to broader systems with more generalized capabilities that are able to respond to different environments and situations and perform a variety of tasks. **While it is hard to predict the exact trajectory of AI advances, what is becoming evident is that the future will witness new forms and structures of collaboration and coordination between humans and AI systems, rather than a future in which AI will replace humans.**

13 Gupta (2022).
14 Browne (2021); Callaway (2020).
15 Gupta (2022).
16 Appen (2022).

PART II: THE DISRUPTIVE IMPACT OF AI ON INTERNATIONAL PEACE AND SECURITY

As the field of AI is evolving rapidly, it is becoming apparent that AI presents unprecedented opportunities to augment human capabilities, particularly in problem-solving and decision-making. However, at the same time, significant ethical, legal, safety and security concerns remain and are coming to the fore as AI systems are increasingly adopted across sectors.[17] These concerns are compounded in the peace and security context, where AI has the potential to transform the conduct of military operations by enabling disruptive increases in efficiency, speed and precision across military uses. While these enhanced capabilities offer benefits for military decision-making and the conduct of military operations, they also present unique risks. Moreover, as an enabling technology, the integration of AI technologies across domains of warfare – from cyber and biological to nuclear, especially in convergence with other powerful dual-use technologies – can have benefits for as well as pose novel risks to international peace and security.

However, advances in AI also present novel opportunities to build peace by offering solutions that can be harnessed for humanitarian purposes – in conflict prevention and peacebuilding processes. More broadly, AI can become an accelerator for the achievement of the United Nations Sustainable Development Goals. AI-based object recognition and predictive analytics are already being harnessed for social good, from disease detection and drug discovery to predicting and limiting the impacts of climate change.[18]

Against this backdrop, through three panels this session examined the disruptive impact of AI on international security. This includes the risks and benefits of uses of AI in military operations and across domains of warfare as well as the opportunities and challenges of harnessing AI technologies for conflict prevention and peacebuilding.

Uses of AI in military operations[19]

AI technologies are developing rapidly and their steady integration into an increasing number of military applications could transform the conduct of military operations by enabling disruptive increases in efficiency, speed and precision. While this may offer benefits to militaries across a range of applications, it also presents unique risks to international peace, security and stability. Against this backdrop, this panel examined how AI developments could potentially transform the conduct of military operations. It discussed the military uses of AI and the associated potential risks, challenges and benefits. It also reflected on how governance frameworks can address the risks and dangers of military applications of AI. While autonomy in weapon systems is indeed one of the most important issues in the context of international

[17] IBM (2022).
[18] Höne (2022).
[19] This section summarizes the discussions that took place during the segment "Panel II: Uses of AI in military operations"; https://www.unidir.org/ID22. This panel was comprised of S. Kate Devitt (Chief Scientist, Trusted Autonomous Systems Defence Cooperative Research Centre and Adjunct Professor QUT Centre for Robotics, Queensland University of Technology), Martin Hagström (Programme Manager, Swedish Defence Research Agency), Margarita Konaev (Deputy Director of Analysis, Center for Security and Emerging Technology) and Kerstin Vignard (Senior Analyst, Johns Hopkins University Applied Physics Lab; Research Scholar, Science Diplomacy & Tech Policy, Institute for Assured Autonomy; Non-resident Senior Fellow, UNIDIR).

peace and security due to the significant legal, safety, security and ethical questions it raises, the discussions drew attention to the wide spectrum of military uses in which militaries are currently looking to leverage AI. These range from AI in military decision-support systems and ISR operations to tactical, operational and strategic planning, military training and logistics. All of these applications could, in varying degrees, have an impact on the use of force.

Often in discussions on military applications of AI, there tends to be a lot of hype surrounding AI technologies and the capabilities they can offer – a perception that AI is like magic fairy dust that can provide solutions to all military challenges. **Nevertheless, as militaries are increasingly experimenting in the use of AI technologies across a range of military systems, processes and practices, there is greater understanding of their current technical limits and the challenges to their adoption at scale. There is also greater understanding of the role of AI in military operations – that AI is not a weapon in itself but an enabler.** But what capabilities is it meant to enable? First is *speed*. The integration of AI technologies through the military decision-making cycle – observe–orient–decide–act (OODA) loop – is meant to give militaries tactical and strategic advantage by dramatically increasing the speed with which military operators and decision makers can execute the OODA loop. Second is *coordination*. Modern warfare is extremely complex. The theatre of operations can extend across different domains involving different forces, systems and decision makers outside the base of operations that need to communicate and coordinate with one another in order to be effective. The integration of AI in military communications and logistics is meant to improve that coordination. Third is *survivability*. Autonomous systems can enhance the ability of militaries to endure and persist in harsh or adversarial environments where human operators may be unable to operate effectively or at all. They can operate in anti-access/area-denial environments, which would help to minimize the number of human operators at risk and, moreover, systems can be made smaller, faster and more agile and thus more combat capable.[20] The fourth is *precision*. AI-enabled image-recognition and object-detection capabilities could help improve military precision in target recognition by being able to quickly analyse large volumes of incoming imagery and video from multiple sensors.[21]

While AI technologies could offer the benefits in military operations outlined above, they also create unique risks. For example, increased speed in military decision-making could result in miscalculation and inadvertent escalation. **Furthermore, current AI technologies are brittle and prone to making errors and being fooled by adversarial spoofing or hacking. As a result, they may ultimately be less accurate and precise than human operators in complex battlefields.** The impetus for integrating AI technologies in military decision-making is the premise that human decision-making can be flawed, especially under pressure in a war situation due to human factors such as biases and emotions. The introduction of data-informed or data-based recommendations and decisions could thus help alleviate human limitations and achieve precision and accuracy. However, AI systems that generate these recommendations can be biased

[20] Morgan et al. (2020).
[21] This para summarises Margarita Konaev's remarks at the panel discussion. Konaev (2022).

in different ways and could reinforce existing human biases at scale (as explained in Part I). Moreover, there may be tension between precision in real-time decision-making and the predictability and understandability of AI systems. Complex higher performing AI systems that can learn in real time and adapt according to changes in dynamic operational environments tend to be less understandable and thus less predictable.[22]

With respect to how AI systems are being used in military operations today, it is important to note that at present their uses are rudimentary and not at scale. But they are perhaps groundbreaking in the sense that these uses have not been attempted before. The most prominent uses of AI technologies today are in ISR operations, where AI systems are being leveraged for gathering, processing and analysis of intelligence data. Other military applications where the integration of AI technologies is being increasingly explored include strengthening cyber defences, conducting as well as mitigating AI-enhanced influence operations, and enhancing combat simulations for military training and planning. However, the ability to use AI capabilities effectively for warfighting – where they would have to be able to adapt to the complicated realities and uncertainties of warfare – is still unproven and, in many ways, untested.[23]

When discussing the possible military applications of AI, it is also important to consider that there are many practical challenges to the adoption of AI in military operations. For one, the military technology acquisition process – from when a decision to acquire a technology is made to when it is fully deployed – can typically span 10–15 years, while technologies like AI are advancing yearly, sometimes monthly. Even new ultra-fast acquisition processes would take at least two years as militaries need to comply with high safety and security standards, train the personnel who would be employing the technology, and develop formal directives, handbooks and manuals to guide the deployment and use of new technologies. One of the key challenges that military organizations are facing today is how to keep the speed of modernization at pace with rapid technological developments.[24] Nevertheless, in the short-to-medium term we can expect to see that the gradual integration of AI technologies in military software, hardware and missions will engender novel forms of human–machine interaction and human–machine teaming – something that we have not encountered before, especially under situations of duress.[25]

As governments around the world are increasingly seeking to harness AI technologies across sectors, including in the military domain, they are developing national AI strategies and even defence-specific AI strategies that define and communicate their vision, values, objectives and intended actions towards the responsible development and use of AI. **To ensure that such high-level strategy documents can translate into the desired operational impact, governments must remain cognizant of three key considerations. First, *AI is all about trade-offs.***

[22] Holland Michel (2020); Konaev (2022).
[23] Konaev (2022).
[24] Hagström (2022).
[25] Konaev (2022).

When you train an algorithm to optimize for one objective, it is not optimizing for another objective. Therefore, it is crucial that actors involved through the AI system lifecycle have clear guidance and methods to consider trade-offs with respect to how to prioritize among different objectives. This involves trade-offs in the operationalization of high-level principles to guide the responsible, ethical and trustworthy research, design, development, deployment and use of AI. These include consideration of how to weigh the operationalization of different principles against each other and how that would have an impact on system performance and how the principles should be put into practice within the constraints of project timelines and budgets. **Second, *AI innovation involves uncertainty*.** While there are scores of examples in recent years of how AI technologies have surprised us with their exceptional performance, there are also sufficient examples that show that the outputs they generate may not always be benign or beneficial. In recognition of the reality that innovation involves uncertainty, it would be prudent for governments to take a risk-based approach towards AI governance that would entail frameworks and mechanisms for risk identification, assessment, mitigation and redress. **Third and most important, *not every nail needs an AI-enabled hammer*.** As governments, particularly their defence sectors, develop strategies for AI adoption, they must put in place processes to determine what problems are good candidates for an AI-enabled solution and which of those solutions are worth pursuing given the state of AI technologies and their likely short-to-medium term trajectories.[26]

Along with AI strategies, **governments around the world are developing governance frameworks with a view to mitigating the risks and dangers of AI applications while harnessing the benefits**, including in the military domain. Similarly, in the civilian sphere, many industry actors that are at the forefront of AI innovation are also adopting principles and guidelines for the responsible research, design, development and use of AI across their operations. **Given that the civilian AI industries work for or with militaries to build AI systems, it is imperative that governments and their militaries take measures to ensure that civilian and military governance frameworks align with respect to military applications of AI technologies.** In this way, those from the AI industry and research community that support defence innovation should comply with national principles, norms or expectations for responsible development and use of AI in the military domain.[27]

While there are overlaps between civilian and military AI governance frameworks with respect to the relevance of certain ethical principles, there are also notable differences. For example, both civilian and military frameworks generally emphasize principles of responsibility, accountability, reliability, bias mitigation and so on. An important difference is that military governance frameworks often focus more on consideration of lawfulness as it concerns compliance with the principles, rules and requirements of international humanitarian law and international human rights law. **Moreover, even though there may be overlaps between civilian and military frameworks with respect to which principles they consider to be essential**

[26] This para summarises Kerstin Vignard's remarks at the panel discussion. Vignard (2022).
[27] Devitt (2022).

for the responsible development and use of AI, the civilian sector and the military may interpret, prioritize or operationalize these principles differently. Thus, to be able to feasibly integrate and deploy AI technologies across military applications such that they are fit for purpose and able to make it through military testing and evaluation processes, militaries will need to clearly define and communicate their expectations and requirements from a safety, security and legal perspective to stakeholders from industry and the research community.[28]

The disruptive impact of AI across domains of warfare[29]

This panel set out to survey the potential benefits and risks of integrating AI technologies across domains of warfare and examine what measures can be taken to mitigate the risks, with a focus on the cyber, biological and nuclear domains.

AI and cyberspace[30]

AI technologies are being increasingly leveraged in various cybersecurity applications, including the detection of and defence against cyber threats. With respect to offensive cyber operations, the picture so far remains largely unclear as to how exactly AI technologies are being used.[31] However, the use of malicious AI-generated synthetic media commonly known as "deepfakes" has emerged in disinformation campaigns.[32] This has raised serious concerns for national and international peace and security.

In cyber defence, AI has proven to be highly effective in providing protection against cyber-attacks ranging from phishing, via spamming to malware attacks.[33] At present, AI is helping organizations to monitor the cybersecurity of systems and customers and to evaluate and address a large number of cyber alerts automatically. This allows cybersecurity analysts to focus on more important and serious incidents and thus avoids the so-called "alert fatigue". **In addition, AI technologies are also deployed to detect potential cyberattacks and infrastructure malfunctions, preventing such incidents from destabilizing national and international peace and security.**

Nevertheless, the use of AI technologies in the cyber domain could potentially increase the risk of malicious cyber activities and enable malicious actors to carry out more targeted attacks at scale with enhanced speed, effectiveness and sophistication. In addition to facilitating cyber-attacks such as distributed denial-of-service (DoS) attacks, AI technologies could also be leveraged for information manipulation, which would serve as a powerful tool in

---

28 Devitt (2022); Vignard (2022).

29 This section summarizes the discussions that took place during the segment "Panel III: The Disruptive Impact of AI Across Domains of Warfare"; https://www.unidir.org/ID22. This panel was comprised of Li Bin (Professor, Department of International Relations, Tsinghua University), Alexander Liskin (Head of Threat Research, Kaspersky), Eleonore Pauwels (Senior Fellow, Global Center on Cooperative Security) and Andrew Reddie (Faculty, University of California, Berkeley).

30 This is based on Alexander Alexander Liskin's remarks at the panel discussion. Liskin (2022).

31 Liskin (2022).

32 Anand and Bianco (2021).

33 Liskin (2022).

warfare. Furthermore, AI technologies, as with all other forms of digital technology, could be potentially hacked – thus increasing the vulnerability of information and communications systems – and there have been real-life examples of such incidents.[34] Lastly, it remains difficult for humans to understand specific results and outputs of complex AI systems, which reduces the trustworthiness of the systems.

AI and biotechnologies[35]

Rapid digital transformation is causing a revolution in biotechnology. In the last two decades, biotechnology has transformed from analogue to digital, with AI as an innovation catalyst.[36] The convergence of AI and biotechnologies has had powerful implications across biotechnology sectors, from bio design and precision medicine to biosecurity. Further, the collaboration between AI and biological engineers has given rise to new applications in functional genomics and proteomics (the large-scale study of proteins), where AI systems can learn to map, analyse, and model and predict the functions of genes and proteins as well as critical interactions between them.[37] Modelling genes and protein functions has transformative benefits for drug discovery, allowing us to understand the impact that a particular pathogen could have on the immune system, whether a virus has high transmissibility between humans and why the genome of certain populations is more susceptible to certain types of infection. As the two fields are rapidly advancing, new technologies and applications continue to emerge at their intersection.

In the area of biosecurity, AI technologies can enhance preparedness and response to large-scale public health emergencies, whether natural, accidental or intentional; facilitate the development of effective medical countermeasures including vaccines, particularly in the event of a public health emergency such as the COVID-19 pandemic;[38] and mitigate or even prevent a biological incident by improving the surveillance and detection of non-natural biological agents and agents that may pose a risk, as well as potential misuses such as illicit gene synthesis. Moreover, modern biotechnologies have become more accessible and connected through AI automation and computing, enabling a wider range of actors to share expertise and accelerate progress in global biotechnology research and development.[39]

---

34 Liskin (2022).
35 This is based on Eleonore Pauwels's remarks at the panel discussion. Pauwels (2022).
36 Pauwels (2021).
37 On functional genomics see NHS Health Education England (2020).
38 Ransbotham, Khodabandeh and Johnson (2021).
39 Anand (2020).

On the other hand, with the adoption of AI technologies in the biotechnology field, almost all aspects of modern biotechnology – from design and experimentation to production – could now be automated and outsourced in decentralized workflows and supply chains.[40] This has facilitated the transfer of dual-use knowledge to a range of non-state actors and has democratized access to the design, development and production of biological agents. While the democratization of biotechnology is driving innovation, it could be exploited by malicious actors for hostile purposes. For example, decentralized technologies for the synthesis, editing and assembly of genes could enable an actor with enough expertise to modify and synthesize a genetic sequence to produce the basis of a toxin or biological agent. Furthermore, recent research has pointed to the potential risk that AI-enabled drug-discovery processes for pharmaceutical purposes could be misused for the design of new biological and chemical agents to cause harm.[41]

AI, nuclear risk and strategic stability[42]

The integration of AI technologies into the nuclear deterrence architecture is currently being researched, developed and in some cases even deployed, particularly for early-warning and ISR systems and for enhancing situational awareness, including the location of enemy nuclear forces. For instance, AI-enabled ISR platforms could be deployed in complex and hostile environments to identify and locate enemy nuclear forces with greater accuracy or through real-time data processing to alert commanders of potential incoming threats or the movements of an adversary's nuclear forces. Moreover, AI systems could help commanders mine and analyse large volumes of intelligence data and in doing so support commanders to more accurately and rapidly anticipate the nature of potential threats or pre-empt a potential strike.[43] **Although it is important to consider that, while applying AI technologies in an early-warning and ISR context may be security maximizing for one state by enhancing its first-strike capabilities, it may be destabilizing for another. This may, in turn, increase nuclear risk and undermine strategic stability, rather than strengthen it.**

Another area that is gaining attention is the use of AI technologies for decision-support in nuclear command and control. There is debate about which decisions require humans in-the-loop and which are appropriate to delegate to machines. However, since current complex AI systems can be unpredictable, difficult to understand and brittle, **there is general agreement among nuclear experts that critical decisions that may have a direct impact on nuclear command and control should be left to humans**. Moreover, one of the key capabilities that AI offers in the military context is speed in decision-making. The risks that increased speed poses (as discussed above) are especially severe in the nuclear context, as swift action and reaction times could cause miscalculations and lead to inadvertent nuclear weapon use.[44]

---

40 Pauwels (2022).
41 Urbina et al. (2022).
42 This is based on Andrew Reddie and Li Bin's remarks at the panel discussion. Reddie (2022); Li (2022).
43 J. Johnson (2022).
44 Li (2022).

Furthermore, while AI cyber defensive tools could enhance the cybersecurity of nuclear command and control systems by detecting and averting cyber intrusions, a motivated adversary could also use malware to hack or fool AI systems.[45]

Lastly, with respect to nuclear disarmament and arms control, AI technologies can be leveraged to support monitoring and verification of states' compliance with nuclear treaty provisions, export control regulations and other international commitments.

Addressing the disruptive impact of AI across domains of warfare

AI technologies will continue to advance in the future, and as such their integration into various domains of warfare will only increase. The key challenge for international peace and security is how to effectively mitigate the potential risks. Through the discussions, four key risk-mitigation measures were highlighted.

First, it is critical to ensure multi-stakeholder engagement in relevant discussions on the convergence of AI with other dual-use technologies, as development and innovations often take place outside the military domain. Transparency and accountability should be promoted among different stakeholders, from the private sector and academia to governments and international organizations, to ensure that AI is researched, designed, developed, deployed and used in a reliable and responsible manner.

Second, as current AI systems can be unpredictable, vulnerable to cyberattacks, brittle and prone to making errors and to adversarial manipulation, **it is essential that military operators can exercise appropriate levels of human oversight, judgement and control over AI-enabled military systems and decision-making processes**. This would minimize the risk of errors and failures that could result in the loss of life or damage critical infrastructure.

Third, the transfer of dual-use knowledge should be properly managed and monitored so as to prevent malicious actors from acquiring the necessary data and technological know-how to cause harm. Further discussion on this issue is urgently needed.

Lastly, testing, monitoring and verification mechanisms should be in place to detect and prevent potential malicious incidents involving the use of AI across different domains. The AI technology itself, with powerful data-processing capacities, could serve as an effective tool to monitor and mitigate potential risks.

[45] J. Johnson (2022).

AI for peace – AI and conflict prevention and peacebuilding[46]

AI technologies could not only transform the conduct of military operations but could equally support important efforts to build and sustain peace. United Nations agencies and humanitarian organizations have been exploring and, in some instances, even applying innovative technologies to address pressing issues in conflict settings around the globe. This has included the use of natural language processing and data analysis.[47] With a focus on harnessing AI solutions for conflict prevention and peacebuilding, this panel examined which AI-enabled tools can be used in this context, how and for what purposes, and discussed the potential challenges of leveraging AI solutions in support of peace efforts.

There are various ways in which AI can potentially play a role in efforts undertaken by peace practitioners. Often, conflicts are driven by divergence of opinions and fragmentation.[48] **AI-powered solutions can help better understand the various views of different communities through media analysis and then to bridge any differences with digital dialogue.** By processing and analysing content on both social media and mainstream media such as newspapers, AI can better inform policymakers of the different needs and concerns within a local context. Furthermore, real-time conversations can be hosted on AI-powered platforms where a large group of individuals can communicate in local dialects and languages, thus increasing inclusivity in the peace process. Furthermore, the analytical capabilities of AI technologies can also aid other aspects of conflict prevention and peacebuilding, including geospatial analysis (such as identifying movements of objects in ceasefire monitoring), detection of emerging trends and patterns, as well as foresight analysis.[49]

In recent years, United Nations agencies have explored and even deployed AI-enabled applications for conflict prevention and peacebuilding around the globe. Inclusivity is an important element for the success of peace processes, and the United Nations Department of Political and Peacebuilding Affairs (DPPA) has started to use AI technologies to facilitate digital dialogues and identify points of agreements in support of peace processes. Notably, since 2019, the DPPA has been partnering with an AI company Remesh to explore AI-enabled approaches to public surveying in the context of conflict resolution and peacebuilding.[50] Leveraging its AI-based platform as a dialogue tool, the DPPA and Remesh have conducted large-scale digital dialogues in local dialects to better understand public perceptions pertaining to conflicts, including in the Syrian Arab Republic, Yemen and Libya, and the United Nations' peace mediation efforts in these conflict settings. The digital platform can enable up to 1000 participants to anonymously engage in simultaneous conversations on a mobile-accessible web interface.[51]

---

[46] This section summarizes the discussions that took place during the segment "Panel IV: AI for Peace – AI and Conflict Prevention and Peacebuilding"; https://www.unidir.org/ID22. This panel was comprised of Paula Hidalgo-Sanchis (Senior Programme Manager, United Nations Global Pulse), Andrew Konya (Founder, Remesh) and Martin Waehlisch (Team Leader, Innovation Cell, United Nations Department of Political and Peacebuilding Affairs).

[47] As demonstrated by the list of projects undertaken by UN DPPA Innovation Cell: https://futuringpeace.org

[48] Waehlisch (2022).

[49] Waehlisch (2022).

[50] Company website: https://www.remesh.ai

[51] Alavi et al. (2022).

Through Remesh's AI-based platform, survey participants are invited to select responses to multiple-choice questions, express their views freely in open-ended questions and assess their level of agreement with selected proposals from other participants. Within a few minutes, the AI-based platform is able to process the received data entries, identify preferred proposals and quantify their representativeness based on demographic segmentation. This can provide a general overview of the priorities, concerns and narratives of particular identity groups, including minority views that tend to get averaged out in public surveys. In 2020, the United Nations Office of the Special Envoy of the Secretary-General for Yemen (OSESGY) also deployed this AI platform as a dialogue tool to hold virtual consultation with over 500 Yemeni citizens, one-third of them women, on the opportunities and challenges of the ongoing peace process.[52]

However, there remain challenges to practical implementation with respect to deploying AI solutions at scale for conflict prevention and peacebuilding. First, there are multiple practical issues on the ground, including Internet connectivity, a lack of technical literacy and the affordability of digital devices.[53] Conflict-stricken populations can often experience power cuts, limited access to resources and even destruction of livelihood, all of which can greatly hamper their abilities to benefit from the use of AI technologies. In addition, common misconceptions surrounding the use of AI technologies can pose just as large an obstacle. For instance, the use of AI technologies to systematically analyse online communication content in conflict-affected areas can be perceived as a mass-surveillance practice, which may undermine confidence in mediators.[54] Furthermore, AI systems can be biased and often there is limited dialogue between the developers and the users.[55] In such instances, potential biases may not be identified and addressed in the research, design and development phases, compromising trust in and the reliability of AI systems. Lastly, along with the recent progress in AI technologies comes the fear of losing jobs, and practitioners might thus become reluctant to adopt the technology in their routine work.

Nevertheless, measures can be taken to enable the effective and ethical use of AI solutions for conflict prevention and peacebuilding. In the discussion, the following were underscored:

First, it is important for all relevant stakeholders, including AI developers, mediators and local practitioners, to be engaged in dialogue, coordination and co-design of AI-enabled systems. At one end of the process, it is imperative that the developers understand the needs of the people who will use the technology in conflict situations and take measures to meet their needs using AI solutions; at the other end, mediators and relevant local actors should equally familiarize themselves with the technology in question and its applications.

52 United Nations DPPA Innovation Cell (2020).
53 Waehlisch (2022).
54 Lindström (2020).
55 Hidalgo-Sanchis (2022).

Second, it is essential that AI solutions are used in compliance with existing legal frameworks and regulations and relevant guidelines such as the recommendation of the United Nations Educational, Scientific and Cultural Organization (UNESCO) on the Ethics of Artificial Intelligence and the United Nations Guidance for Effective Mediation.[56]

Third, data protection should be ensured, including by using AI solutions in accordance with good practices such as the United Nations Principles on Personal Data Protection and Privacy.[57] In public surveying, anonymity of the participants should be guaranteed when required and information on, among other things, how the data will be collected, processed and used should be transparent.

Fourth, AI-enabled tools should be sufficiently tested before deployment in order to minimize risks and ensure that the technology will be used in a responsible and ethical manner in the sensitive context of conflict prevention and peacebuilding.

Lastly, local contexts and expertise should be included in the research, design, development, deployment and use of AI systems. This is key to ensuring that the use of AI solutions for conflict prevention and peacebuilding are in harmony with local norms and practice.

[56] UNESCO (2021); United Nations Department of Political Affairs (2012).
[57] United Nations HLCM (2018).

2022 Innovations Dialogue participants entering the auditorium of the Ford Foundation Center for Social Justice in New York

PART III: TOWARDS RESPONSIBLE AI[58]

While it is apparent that AI technologies can augment human capabilities, they present significant ethical, legal, safety and security concerns. These range from issues related, but not limited, to transparency, reliability, predictability, understandability, accountability, bias and discrimination, and technical robustness.[59] Thus, stakeholders involved in different stages of the AI system lifecycle share the goal that AI systems be researched, designed, developed, deployed and used in a responsible manner and in accordance with legal requirements and ethical values. **To ensure this, governments, intergovernmental organizations, private sector entities and members of civil society are developing normative instruments such as principles and standards to guide the AI system lifecycle. This approach to AI governance is broadly known as Responsible AI, or ethical or trustworthy AI.** However, the RAI approach is a young and evolving field of research and practice. More work needs to be done to understand what RAI entails, how it can be put into practice across critical sectors and, particularly for our purposes, how it relates to international peace and security challenges.

Against this backdrop, through two panels, this session unpacked what Responsible AI is and the elements that comprise the RAI toolbox. It further examined how RAI principles and tools are and could be put into practice. Finally, the session reflected on the importance of building a culture of Responsible AI.

What is Responsible AI?

There is no single definition for Responsible AI. What RAI should exactly entail can be context-specific and different stakeholders may therefore define it differently. Nonetheless, at the core, there are growing common views on the cornerstones of RAI. **Essentially, RAI can be understood as a principles-based, socio-technical approach to the research, design, development, deployment, use, maintenance and governance of AI systems across sectors that is conscious of and considers the effects (both positive and negative) that such systems may have on individuals, communities and society at large.**[60] In that, the goal of RAI is to ensure that AI systems and their outcomes are safe, secure, fit for the purpose for which they were developed, reliable, accurate, non-discriminatory, transparent, explainable, lawful, ethical and respecting of people's fundamental rights.

---

[58] This part summarizes the discussions that took place during the segments "Panel V: Unpacking and operationalizing the responsible AI toolbox" and "Panel VI: Building a culture of responsible AI: A shared responsibility"; https://www.unidir.org/ID22. Panel V was comprised of Ashley Casovan (Executive Director, Responsible AI Institute), Rebecca Finlay (Chief Executive Officer, Partnership on AI), Emma Ruttkamp-Bloem (Professor, Department of Philosophy, University of Pretoria, Centre for AI Research, Council for Scientific and Industrial Research), Sonali Sanghrajka (Co-founder and Chief Commercial Officer, Kosa.ai), Eugenio Vargas Garcia (Tech Diplomat, Brazilian Consulate General in San Francisco), Daniel Kluttz (Director of Sensitive Uses at the Office of Responsible AI, Microsoft), Diane Staheli (Responsible AI Chief at the Chief Digital and Artificial Intelligence Office, US Department of Defense) and Alice Xiang (Global Head of AI Ethics, Sony). Panel VI was comprised of Sumaya H. Al Hajeri (AI Expert, UAE AI Expert Group), Adedeji Ebo (Director and Deputy to the High Representative, UN Office for Disarmament Affairs), Marek Havrda (Deputy Minister for European Affairs, Government of Czechia) and Catherine Régis (Full professor of Law and Associate Vice-President, University of Montreal and Co-chair of the Working Group on Responsible AI of the Global Partnership on AI).

[59] There are recent real-world instances where AI systems in use have made consequential mistakes – from a self-driving car killing a pedestrian to perpetuated or newly developed biases in AI-based recruitment tools, as demonstrated in McKendrick and Thurai (2022).

[60] Klutz (2022).

AI systems and the effects they may have are a result of many compounding decisions made by those who research, design, develop, deploy, use and maintain the systems. **Through ethical principles or guidelines and a combination of tools such as (but not limited to) testing standards, risk-assessment frameworks, conformity-assessment schemes, accountability checks and employment guidance, the RAI approach can proactively ensure that decisions made through the AI system's lifecycle result in intended outcomes.** RAI is therefore not just about understanding how AI systems work, but also about understanding why they produce the outcomes or decisions that they do and addressing how those decisions or outcomes could have an impact on individuals, communities and society. **This approach therefore helps to prevent unethical or irresponsible applications of AI technologies and consequently builds trust in AI systems.** The trust in turn is an enabler for the rapid adoption and deployment of AI systems. The RAI approach is also flexible – the tools and methodologies are customizable and tailorable, taking into account the nature, risk level, scope and scale of the specific use case and the situation of use. Furthermore, RAI is not static. Even after an AI system and its use are determined to be responsible, it requires continuous oversight as technology evolves or the context of use of the AI system changes. Therefore, different tools developed to ensure that AI systems are responsible and trustworthy also require constant review and may need to be updated, particularly in accordance with the experience of implementing them through the AI system lifecycle. And, since RAI is a lifecycle approach towards managing risks and preventing possible harms while facilitating responsible use, it is a multi-stakeholder and multidisciplinary endeavour. **Stakeholders involved in and concerned with every stage of the AI system lifecycle – from ideation to use and maintenance – have a shared role to play in RAI and must continuously and systematically engage with one another to address RAI issues through the lifecycle.**

How can Responsible AI be operationalized?

RAI efforts usually begin with the adoption of broad AI principles or guidelines that encompass the technical, legal and ethical requirements that AI systems should meet in order to be responsible and trustworthy. Many governments, industry stakeholders, intergovernmental organizations and civil society actors are adopting such principles to provide direction to the entire lifecycle of an AI system. In the particular case of governments, it is important to note that, while many have so far taken a sector-agnostic approach to AI principles, others are developing principles tailored to specific sectors as necessary, including the defence sector. At the international level, in 2021 UNESCO notably adopted the first globally accepted (i.e. adopted by its 193 member states) normative instrument on the Ethics of AI in the form of a recommendation. The recommendation provides a framework of principles, values and actions to guide states in the establishment of their own AI governance frameworks as well as the actions of individuals, communities, institutions and private sector companies to ensure ethics and legal considerations are embedded in all stages of the AI system lifecycle.[61] Through inclusive and multidisciplinary consultations that took into consideration different cultural understandings and approaches to certain kinds of ethical values as well as the circumstances and priorities of each member state, UNESCO's recommendation was able to elaborate a bottom line that should be respected in relation to the lawful and ethical development and use of AI.[62]

[61] UNESCO (2021).
[62] Ruttkamp-Bloem (2022).

Committing to principles is, however, not sufficient. **To achieve responsible and trustworthy AI, these broad principles need to translate into practice, and this is a complex task.** For one, while all principles are desirable and would ideally work in harmony, in practice there may be trade-offs between different principles which would require contextual assessments, such as the possible tension between understandability and accuracy of complex AI systems. Moreover, those creating, using and affected by AI systems may interpret and optimize principles differently in accordance with their diverse needs. **Therefore, beyond the commitment to principles, governments and organizations creating or using AI, at their own levels, should develop detailed practical guidance for AI actors involved in the AI system lifecycle and put in place tools, processes, and governance structures and mechanisms for the operationalization of AI principles.**[63] Some governments are already taking action to this end. For example, the US Department of Defense (DoD) adopted Ethics Principles for AI in 2020.[64] Subsequently, in 2021, it put out a RAI Strategy and Implementation Pathway that defines and communicates the DoD's strategic approach to the operationalization of the Ethics Principles.[65] This includes the outlining of over 60 lines of effort that the DoD will take across the following six tenets to realize its vision for RAI:

(1) *RAI governance*, which concerns ensuring disciplined governance structures and processes for oversight and accountability and clearly articulating DoD guidelines and polices on RAI
(2) *Warfighter trust*, which is to be ensured by providing education and training and by establishing a framework for test, evaluation, verification and validation that integrates real-time monitoring, algorithm confidence metrics and user feedback
(3) *AI product and acquisition lifecycle*, which entails developing tools, policies, processes, systems and guidance for ensuring RAI implementation for an AI product throughout the acquisition lifecycle through systems engineering and risk-management
(4) *Requirements validation*, which concerns incorporating RAI into all applicable AI requirements, including joint performance requirements
(5) *The RAI ecosystem*, which concerns building a robust national and global RAI ecosystem
(6) *The AI workforce*, which concerns building, training, equipping and retaining *a RAI*-ready workforce.

For each line of effort, the DoD's Pathway also specifies which organization is responsible for executing the actions. Furthermore, it provides a non-exhaustive list of tools that can be utilized by the DoD to implement the Pathway, such as templates for AI project management, a toolkit for AI testing and evaluation (including tools and technologies to detect both adversarial attacks on AI and natural degradation of AI systems performance), and a toolkit for acquisition including RAI-related evaluation criteria and standard AI contract language.[66]

---

[63] An AI actor can be defined as any actor involved in at least one stage of the AI system life cycle. It can refer to both natural and legal persons, such as researchers, programmers, engineers, data scientists, end-users, business enterprises, universities and public and private entities, among others.
[64] US Department of Defense (2020).
[65] US Department of Defense (2022).
[66] US Department of Defense (2022).

Similarly, many private companies developing or using AI technologies are adopting RAI principles or ethics guidelines and putting in place tools and mechanisms for the implementation of RAI across operations. For example, Sony adopted its AI Ethics Guidelines in 2018, which include principles like fairness, transparency, stakeholder engagement, trustworthiness and building AI that betters society.[67] Sony then established an AI Ethics Committee, which examines high-stakes AI use cases to ensure compliance with the ethics guidelines. Subsequently, Sony also set up an AI Ethics Office, which is responsible for operationalizing its AI Ethics Guidelines across all business units and evaluating both high- and low-risk AI use cases. In addition, Sony has established research labs that focus on AI ethics issues and on developing benchmarks, diagnostic tools and techniques to build ethical and responsible AI. Notably, it has also made conducting an AI ethics assessment a mandatory part of the quality-management system for its electronics products. For this, Sony uses an AI ethics-by-design approach that entails checking for potential AI ethics issues at every stage of the AI development lifecycle, from research and design to development and deployment. This is important because certain issues (including those related to design and the data being used) cannot be adequately addressed at the end of the process, and other ethics-relevant issues (including the exact fairness properties of an AI system) are easier to evaluate at later stages of the lifecycle.[68]

Likewise, Microsoft developed its RAI principles in 2018.[69] These include considerations of fairness, accountability, transparency, reliability, safety, inclusiveness, privacy and security. To operationalize the principles, Microsoft has adopted a hub-and-spoke model.[70] The "hub" includes Microsoft's Office of Responsible AI, which is the governance, policy and compliance engine for RAI. In addition to this is the Aether Committee, comprised of Microsoft's top scientific and engineering talent, who provide subject-matter expertise on the state-of-the-art and emerging trends in the context of the implementation of Microsoft's RAI principles. Lastly, the hub includes a Responsible AI Strategy in Engineering group that helps the Microsoft engineering groups to implement RAI processes through systems and tools. The "spoke" aspect of the governance approach includes a community of RAI champions, who are appointed by company leadership to sit in engineering and sales teams and raise awareness and support the implementation of Microsoft's RAI governance approach, policies and tools.[71] With respect to guidance for the implementation of the RAI principles, Microsoft has developed a Responsible AI Standard that is now publicly available. The Standard serves as an internal playbook of guidelines and requirements that all engineering teams must meet in the development and deployment of AI systems. For each requirement, the Standard enables the engineers to conduct impact assessments of the systems they are developing and deploying and encourages them to articulate and document why the AI system is fit for the intended use and what the risks and benefits could be in context of their stakeholders. It also sets out RAI implementation tools and practices that the teams can draw upon.[72]

---

67 Sony (2018)..
68 Xiang (2022); Sony (n.d.).
69 Microsoft (n.d.).
70 Crampton (2021); Klutz (2022).
71 Crampton (2021).
72 Crampton (2021); Klutz (2022).

Notably, Microsoft has a specialized process for an additional level of review and oversight over sensitive or high-impact use cases. These include cases where AI systems in their intended use or through their misuse could pose a risk to human rights, infringe upon people's legal positions or life opportunities, or cause physical or psychological injury. For such sensitive uses, Microsoft has set up a reporting function where an engineering team that is developing or deploying an AI system that in its end use could meet one of the triggers must report it to the sensitive uses team in the RAI Office, which will then work with the engineering team to implement the RAI Standard.[73]

There are also many start-ups emerging that offer consulting services on RAI issues or technical tools to address issues surrounding RAI, such as fairness and explainability. For example, Kosa.ai offers a range of plug-and-play software solutions that can be integrated with any data sources or model frames to help organizations that are developing or using AI to operationalize RAI principles and build trust in their AI-enabled products.[74] The tools that it offers are designed for easy use by both technical and non-technical AI actors and cater to both the pre-market and post-market environment. Pre-market tools are to assure the safety, robustness, fairness and effectiveness of systems that are still under development or nearly ready for the market. Kosa.ai provides post-market tools for surveillance, monitoring and explainability of AI systems or AI-enabled products that are already in production or on the market.[75]

At the international level, as an outcome of the Ethics of AI recommendation, UNESCO is currently developing a tool for carrying out ethical impact assessments (EIAs) in consultation with a multidisciplinary and multicultural advisory group of experts. The aim of the EIA tool is to help UNESCO member states implement the recommendation.[76] It is envisaged to help AI actors ensure alignment with the ethical principles provided in the recommendation. To this end, it will identify, monitor and assess the benefits and risks of AI systems as well as anticipate consequences, mitigate risks, avoid harmful outcomes and address societal challenges.[77]

An essential element of bridging the gap between principles and practice is for an organization to produce accurate and useful documentation for every AI system or AI-enabled product that it develops. Documentation methodologies and practices can improve understanding of consequential decisions made through the AI system lifecycle, including how algorithms are being developed or how and for what reason the data is being used, and their potential effects.[78] Asking the right questions for each type of AI system at the right time in the development process would not only assist in anticipating and mitigating potential RAI issues but would also improve transparency. To this end, the Partnership on AI through its multistakeholder "ABOUT ML" initiative is bringing together a diverse range of expertise to develop, test and implement AI system documentation practices at scale.[79]

---

73 Klutz (2022).
74 Company website: http://www.kosa.ai/about
75 Sanghrajka (2022).
76 UNESCO (2022).
77 UNESCO (2022).
78 Finlay (2022).
79 Partnership on AI (n.d.); Custis (2021).

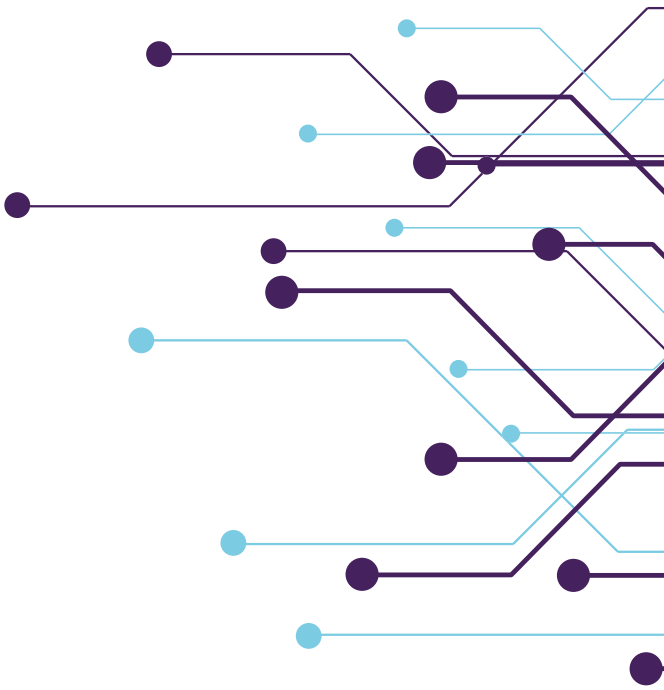
An important consideration in relation to operationalization of RAI is that AI is not a monolithic technology. Therefore, the tools employed for RAI operationalization need to be contextualized for specific types of AI system, their uses and what their different implications could be.[80] This could be addressed through the development of RAI standards that would help to build a common understanding of the impacts, risks and harms of specific types of AI system and their uses. These standards would also allow comparability and repeatability with respect to understanding the implications of different AI systems and how they can be addressed. Each AI system could be objectively evaluated based on such standards and be given certification that would signify that the AI system or AI-enabled product conforms with standards and thus is trustworthy.[81] To this end, the Responsible Artificial Intelligence Institute has developed an independent and accredited conformity assessment tool that provides assurance that AI systems as well as organizational practices are aligned with requirements of RAI specifications, best practices, regulations and standards.[82] The tool can be used for self-assessment, an independently delivered assessment or a certification delivered by accredited auditors.[83]

Building a culture of Responsible AI

Scaling up RAI practices and realizing the adoption of responsible and trusted AI systems ultimately requires cultivating and sustaining a culture of RAI. In such a culture, RAI-related considerations and values are instilled in the organizational culture and viewed as an integral and enabling part of AI development, rather than barriers to it, at both the systemic and individual levels. This can be achieved through RAI education, awareness raising and upskilling of AI actors and by continuous community-wide multidisciplinary consultations, communication and collaboration, particularly between the policy and technical communities with respect to conducting impact assessments of AI developments. RAI considerations and values need to be at the centre of AI research and design thinking. In that, in the AI research and design phases itself AI researchers should be sensitized and skilled to consider the possible ethical, legal, safety and security impacts of their work and to collaborate with AI ethics and legal experts as necessary. In this regard, universities can play a critical role by providing RAI-oriented education, skills and training to young AI scientists in the early stages of their career. Likewise, the RAI policy community should engage with the AI research community to build understanding of the AI research and development process to be able to develop RAI policies and tools that are technically implementable and viable.

[80] Casovan (2022).
[81] Casovan (2022).
[82] Responsible Artificial Intelligence Institute (n.d.).
[83] Website of Responsible Artificial Intelligence Institute: https://www.responsible.ai

Another critical aspect of building and sustaining an RAI culture is strong leadership at an organizational level and demonstration of good practice by influential actors, in both the public and private sectors, through which norms of responsible behaviour can emerge. Governments can play a critical role here by demonstrating good practice through ensuring transparency and coherence in their national policies and by putting in place structures and mechanisms to operationalize them. Good practice in the public sector can set baseline standards, norms and expectations of responsible behaviour.[84] At the organizational level, leaders must articulate, communicate and socialize a vision for a firm commitment to RAI; demonstrate that it is a top management priority because it enables AI adoption rather than restricts it; and develop clear pathways for RAI adoption through a consultative process involving internal and external stakeholders.



[84] Vignard (2022).

REFERENCE LIST

Alavi, Daanish Masood et al. 2022. "Using artificial intelligence for peacebuilding". *Journal of Peacebuilding & Development*, vol. 17, no. 2 (August 2022): 239–243. https://doi.org/10.1177/15423166221102757

Anand, Alisha. 2020. "The 2020 Innovations Dialogue conference report". Geneva, Switzerland: United Nations Institute for Disarmament Research. 1 December. As of 1 February 2023: https://unidir.org/publication/2020-innovations-dialogue-conference-report

Anand, Alisha and Belen Bianco. 2021. "The 2021 Innovations Dialogue conference report". Geneva, Switzerland: United Nations Institute for Disarmament Research. 22 December. As of 1 February 2023: https://unidir.org/publication/2021-innovations-dialogue-conference-report

Appen. 2022. "Autonomous vehicles: One of AI's most challenging tasks". 19 July. As of 1 February 2023: https://appen.com/blog/autonomous-vehicles-the-most-challenging-task-in-ai

Browne, Grace. 2021. "DeepMind's AI has finally shown how useful it can be". WIRED. 22 July. As of 1 February 2023: https://www.wired.co.uk/article/deepmind-protein-folding-database

Callaway, Ewen. 2020. "'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures". *Nature*. 30 November. As of 1 February 2023: https://www.nature.com/articles/d41586-020-03348-4

Casovan, Ashley. 2022. "Panel V: Unpacking and operationalizing the Responsible AI toolbox", panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

Crampton, Natasha. 2021. "The building blocks of Microsoft's responsible AI program". Microsoft on the Issues blog. 19 January. As of 1 February 2023: https://blogs.microsoft.com/on-the-issues/2021/01/19/microsoft-responsible-ai-program

Custis, Christine. 2021. "Operationalizing AI ethics through documentation: ABOUT ML in 2021 and beyond". Partnership on AI Blog. 14 April. As of 1 February 2023: https://partnershiponai.org/about-ml-2021

Devitt, S. Kate. 2022. "Panel II: Uses of AI in military operations", panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

Finlay, Rebecca. 2022. "Panel V: Unpacking and operationalizing the responsible AI toolbox", panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

Friedman, Batya and Helen Nissenbaum. 1996. "Bias in computer systems". *ACM Transactions on Information Systems (TOIS)*, vol. 14, no. 3 (July 1996): 330–347. https://doi.org/10.1145/230538.230561

Gupta, Abhishek. 2022. “Panel I: What even is, AI? – The state of play and the future of AI”, panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

Hagström, Martin. 2022. “Panel II: Uses of AI in military operations”, panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

Hidalgo-Sanchis, Paula. 2022. “Panel IV: AI for peace – AI and conflict prevention and peace-building”, panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

Holland Michel, Arthur. 2020. “The black box, unlocked: Predictability and understandability in military AI”. Geneva, Switzerland: United Nations Institute for Disarmament Research. 22 September. https://doi.org/10.37559/SecTec/20/AI1

Höne, Katharina. 2022. “An algorithm for peace? AI in international peace mediation”. Israel Public Policy Institute Commentary. 23 May. As of 1 February 2023: https://www.ippi.org.il/an-algorithm-for-peace-ai-in-international-peace-mediation

IBM. 2022. “IBM Global AI Adoption Index 2022”. May. As of 1 February 2023: https://www.ibm.com/downloads/cas/GVAGA3JP

Johnson, James. 2022. “AI, autonomy, and the risk of nuclear war”. War on the Rocks. 29 July. As of 1 February 2023: https://warontherocks.com/2022/07/ai-autonomy-and-the-risk-of-nuclear-war

Johnson, Khari. 2022. “DALL-E 2 creates incredible images—and biased ones you don’t see”. WIRED. 5 May. As of 1 February 2023: https://www.wired.com/story/dall-e-2-ai-text-image-bias-social-media

Klutz, Daniel. 2022. “Panel V: Unpacking and operationalizing the Responsible AI toolbox”, panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

Konaev, Margarita. 2022. “Panel II: Uses of AI in military operations”, panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

Li, Bin. 2022. “Panel III: The disruptive impact of AI across domains of warfare”, panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

Lindström, Marta. 2020. “Mediation perspectives: Artificial intelligence in conflict resolution”. ETH CSS Blog. 7 February. As of 1 February 2023: https://isnblog.ethz.ch/css-blog/mediation-perspectives-artificial-intelligence-in-conflict-resolution

Liskin, Alexander. 2022. “Panel III: The disruptive impact of AI across domains of warfare”, panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

McKendrick, Joe and Andy Thurai. 2022. “AI isn’t ready to make unsupervised decisions”. *Harvard Business Review*. 15 September. As of 1 February 2023: https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions

Microsoft. n.d. “Microsoft responsible AI principles”. As of 1 February 2023: https://www.microsoft.com/en-us/ai/our-approach?activetab=pivot1:primaryr5

Morgan, Forrest E. et al. 2020. “Military applications of artificial intelligence: Ethical concerns in an uncertain world”. Santa Monica, CA: RAND Corporation. As of 1 February 2023: https://www.rand.org/pubs/research_reports/RR3139-1.html

NHS Health Education England. 2020. “What is functional genomics?” Genomics Education Programme. 29 January. As of 1 February 2023: https://www.genomicseducation.hee.nhs.uk/blog/what-is-functional-genomics

Partnership on AI. n.d. “About ML”. As of 1 February 2023: https://partnershiponai.org/workstream/about-ml

Pauwels, Eleonore. 2021. “Cyber-biosecurity: How to protect biotechnology from adversarial AI attacks”. Hybrid CoE Strategic Analysis no. 26 (May 2021). As of 1 February 2023: https://www.hybridcoe.fi/wp-content/uploads/2021/05/20210503_Hybrid_CoE_Strategic_Analysis_26_Cyber_biosecurity_WEB.pdf

Pauwels, Eleonore. 2022. “Panel III: The disruptive impact of AI across domains of warfare”, panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

Ransbotham, Sam, Shervin Khodabandeh and Dave Johnson. 2021. “AI and the COVID-19 vaccine: Moderna’s Dave Johnson”. *Me, Myself, and AI* (Podcast). 13 July. As of 1 February 2023: https://sloanreview.mit.edu/audio/ai-and-the-covid-19-vaccine-modernas-dave-johnson

Reddie, Andrew. 2022. “Panel III: The disruptive impact of AI across domains of warfare”, panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

Responsible Artificial Intelligence Institute. n.d. “How we help”. As of 1 February 2023: https://www.responsible.ai/how-we-help

Ruttkamp-Bloem, Emma. 2022. “Panel V: Unpacking and operationalizing the Responsible AI toolbox”, panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

Sanghrajka, Sonali. 2022. “Panel V: Unpacking and operationalizing the Responsible AI toolbox”, panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

Schwartz, Oscar. 2019. “In 2016, Microsoft’s racist chatbot revealed the dangers of online conversation”. IEEE Spectrum. 25 November. As of 1 February 2023: https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

Sony (n.d.). “Responsible AI utilization and R&D”. As of 1 February 2023: https://www.sony.com/en/SonyInfo/sony_ai/guidelines.html

Sony. (2018). “Sony Group AI Ethics Guidelines”. 25 September. As of 1 February 2023: https://www.sony.com/en/SonyInfo/csr_report/humanrights/AI_Engagement_within_Sony_Group.pdf

Stoyanovich, Julia. 2022. “Panel I: What even is, AI? – The state of play and the future of AI”, panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

United Nations Department of Political Affairs. 2012. “United Nations Guidance for Effective Mediation”. July. As of 1 February 2023: https://peacemaker.un.org/resources/mediation-guidance

United Nations Department of Political and Peacebuilding Affairs (DPPA) Innovation Cell. 2020. “AI & Peacemaking”. As of 1 February 2023: https://futuringpeace.org/ai-for-peacemaking.html

United Nations Educational, Scientific and Cultural Organization (UNESCO). 2021. “Recommendation on the Ethics of Artificial Intelligence”. 23 November. As of 1 February 2023: https://unesdoc.unesco.org/ark:/48223/pf0000381137

United Nations Educational, Scientific and Cultural Organization (UNESCO). 2022. “Implementation of the recommendation on the Ethics of Artificial Intelligence (AI)”. 7 October. As of 1 February 2023: https://unesdoc.unesco.org/ark:/48223/pf0000382931

United Nations General Assembly (UNGA). 2018a. “Current developments in science and technology and their potential impact on international security and disarmament efforts”. UN document A/73/177, 17 July.

—. 2018b. “Role of science and technology in the context of international security and disarmament”. UN document A/RES/73/32, 11 December.

—. 2021. “Current Developments in science and technology and their potential impact on international security and disarmament efforts”. UN document A/76/182, 19 July.

United Nations High-Level Committee on Management (HLCM). 2018. "Personal data protection and privacy principles". 11 October. As of 1 February 2023: https://unsceb.org/personal-data-protection-and-privacy-principles

United Nations Office for Disarmament Affairs. 2018. "Securing our common future: An agenda for disarmament". May. As of 1 February 2023: https://doi.org/10.18356/80210262-en

United States Department of Defense. 2020. "DOD adopts ethical principles for artificial intelligence". 24 February. As of 1 February 2023: https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence

United States Department of Defense. 2022. "Responsible artificial intelligence strategy and implementation pathway". June. As of 1 February 2023: https://media.defense.gov/2022/Jun/22/2003022604/-1/-1/0/Department-of-Defense-Responsible-Artificial-Intelligence-Strategy-and-Implementation-Pathway.PDF

Urbina, Fabio et al. 2022. "Dual use of artificial-intelligence-powered drug discovery". *Nature Machine Intelligence*, vol. 4, 189–191 (7 March). https://doi.org/10.1038/s42256-022-00465-9

Vignard, Kerstin. 2022. "Panel II: Uses of AI in military operations", panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

Von Laufenberg, Roger. 2020. "Bias and discrimination in algorithms – Where do they go wrong?". VICESSE. 29 June. As of 1 February 2023: https://www.vicesse.eu/blog/2020/6/29/bias-and-discrimination-in-algorithms-where-do-they-go-wrong

Waehlisch, Martin. 2022. "Panel IV: AI for peace – AI and conflict prevention and peacebuilding", panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

Xiang, Alice. 2022. "Panel V: Unpacking and operationalizing the Responsible AI toolbox", panel discussion, 2022 Innovations Dialogue, New York, 20 October 2022: https://www.unidir.org/ID22

UNIDIR
UNITED NATIONS INSTITUTE FOR DISARMAMENT RESEARCH

#ID22

the 2022
innovations dialogue.

AI DISRUPTION, PEACE & SECURITY

20 OCTOBER 2022 • 08:30 – 18:00 EDT • NEW YORK & ONLINE

CONFERENCE OPENING — 08:30 - 08:40

Robin Geiß — UN Institute for Disarmament Research

SESSION 1 • DECODING AI – THE STATE OF PLAY AND THE FUTURE OF AI

PANEL I What even is, AI? – The State of Play and the Future of AI — 08:40 - 10:00

This session will provide a foundational understanding of the concept of AI, the state of play of AI technologies and their most important functionalities. It will also reflect on some of the current obstacles to and opportunities for advancement and where AI is headed in the future.

FEATURING

Abhishek Gupta	Montreal AI Ethics Institute and Boston Consulting Group
Jason Lin	Stanford Existential Risk Initiative, Lyft Self-Driving and Google X
Julia Stoyanovich	New York University

MODERATED BY

Ioana Puscas	UN Institute for Disarmament Research

COFFEE BREAK

SESSION 2 • THE RISKS AND REWARDS OF AI FOR INTERNATIONAL PEACE AND SECURITY

PANEL II Uses of AI in Military Operations — 10:15 - 11:20

This panel will discuss how AI innovations could transform military operations. It will examine the potential risks and benefits of the use of AI as a decision-support tool more broadly, taking the discussions beyond autonomous weapons which have been the focus of multilateral discussions. Particularly, this session will discuss the uses of AI technologies in command-and-control systems, intelligence, surveillance, and reconnaissance (ISR) activities and military planning and logistics.

FEATURING

S. Kate Devitt	Trusted Autonomous Systems Defence CRC and Queensland University of Technology
Martin Hagström	Swedish Defence Research Agency
Margarita Konaev	Center for Security and Emerging Technology
Kerstin Vignard	Johns Hopkins University Applied Physics Lab; Institute for Assured Autonomy and UNIDIR

MODERATED BY

Alisha Anand	UN Institute for Disarmament Research

LUNCH BREAK

PANEL III The Disruptive Impact of AI Across Domains of Warfare — 12:20 - 13:40

This panel will survey specific potential risks and benefits of integrating AI technologies as an enabler of autonomy across domains of warfare, especially in its convergence with other powerful dual-use technologies. Particularly, this panel will focus on the following topics: AI and Cyberspace; AI and Biotechnologies; AI, Nuclear Risk and Strategic Stability; and AI and Outer Space.

FEATURING

Li Bin	Department of International Relations, Tsinghua University
Alexander Liskin	Kaspersky
Eleonore Pauwels	Global Center on Cooperative Security
Andrew Reddie	University of California, Berkeley

MODERATED BY

Beyza Unal UN Office for Disarmament Affairs

PANEL IV AI for Peace – AI and Conflict Prevention and Peacebuilding **13:40 - 14:45**

This panel will focus on harnessing AI solutions for conflict prevention and peacebuilding. It will examine which, how and for what purposes AI-enabled tools can be used in this context. Further, it will discuss the potential risks and challenges of leveraging AI solutions for conflict prevention and peacebuilding.

FEATURING

Paula Hidalgo-Sanchis UN Global Pulse
Andrew Konya Remesh
Martin Waehlisch Innovation Cell, UN Department of Political and Peacebuilding Affairs

MODERATED BY

Sarah Grand-Clément UN Institute for Disarmament Research

COFFEE BREAK

SESSION 3 • TOWARDS RESPONSIBLE AI

PANEL V Unpacking and Operationalizing the Responsible AI Toolbox **15:10 - 16:50**

This panel will unpack what Responsible AI is and the elements that comprise the Responsible AI toolbox. It will also examine how the Responsible AI toolbox can be put in practice. What are the best practices? What are the gaps and challenges and how should they be addressed?

UNPACKING THE RESPONSIBLE AI TOOLBOX

Ashley Casovan Responsible AI Institute
Rebecca Finlay Partnership on AI
Emma Ruttkamp-Bloem University of Pretoria and Council for Scientific and Industrial Research
Sonali Sanghrajka Kosa.ai

OPERATIONALIZING THE RESPONSIBLE AI TOOLBOX

Eugenio Vargas Garcia Brazilian Consulate General in San Francisco
Daniel Kluttz Microsoft
Diane Staheli Chief Digital and Artificial Intelligence Office, Department of Defense, USA
Alice Xiang Sony

MODERATED BY

Giacomo Persi Paoli UN Institute for Disarmament Research

PANEL VI Building a Culture of Responsible AI: A Shared Responsibility **16:50 - 17:55**

This concluding panel will discuss the shared roles and responsibilities of key stakeholders – governments, industry, universities, civil society, and the UN – with respect to fostering a culture of Responsible AI design, development, deployment, and use. How can a Responsible AI culture be created and sustained? Who is responsible for Responsible AI? How can we build synergies between bottom-up and top-down approaches to Responsible AI?

FEATURING

Sumaya H. Al Hajeri UAE AI Expert Group
Adedeji Ebo UN Office for Disarmament Affairs
Marek Havrda Government of the Czech Republic
Catherine Régis University of Montreal and Global Partnership on AI

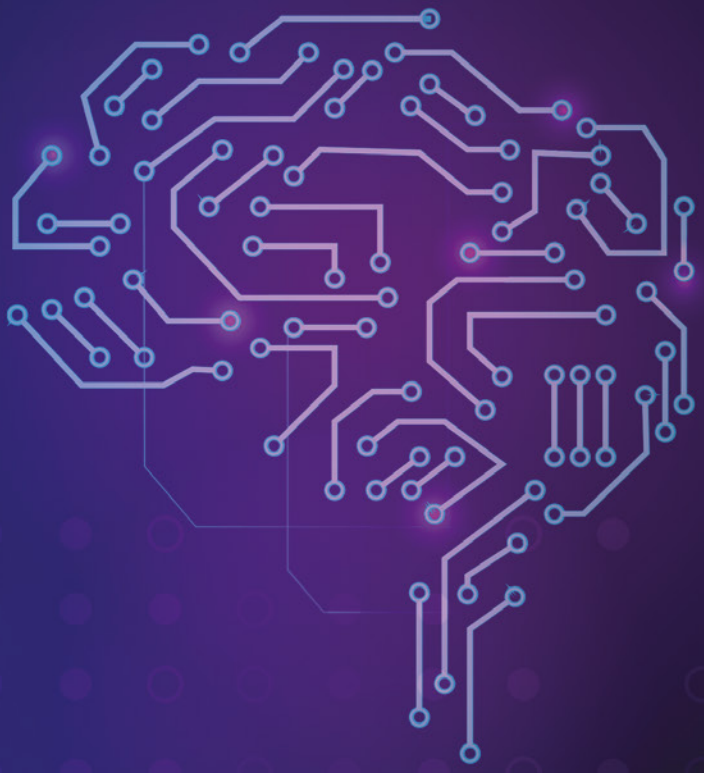
MODERATED BY

Robin Geiß UN Institute for Disarmament Research

CONFERENCE CLOSING **17:55 - 18:00**

Robin Geiß UN Institute for Disarmament Research

RECEPTION *SPONSORED BY THE KINGDOM OF THE NETHERLANDS* **18:00 - 20:00**



UNIDIR UNITED NATIONS INSTITUTE FOR DISARMAMENT RESEARCH

@unidirgeneva @UNIDIR un_disarmresearch