UNIDIR

Exploring the AI–ICT Security Nexus

1 . Introduction

As governments, businesses, and societies grow more digitally interconnected, cyber resilience and cybersecurity strategies have become pivotal issues in safeguarding national and global stability. Artificial intelligence's (AI) application in the information and communication technologies (ICT) domain is reshaping the landscape of both offensive and defensive cybersecurity, providing enhanced capabilities to malicious actors while simultaneously offering unprecedented tools to defenders.

In the ongoing Open-ended Working Group on security of and in the use of information and communications technologies 2021–2025 (OEWG), States are increasingly expressing concerns over threats coming from AI-enabled malicious ICT activities. In the last Annual Progress Report (APR) adopted in July 2024, AI was specifically mentioned in the Existing and Potential Threats section, where States noted that AI (as well as other emerging technologies) "could potentially have implications for the use of ICTs in the context of international security by creating new vectors and vulnerabilities in the ICT space".[1]

However, to support a more concrete examination of the impact of AI, both positive and negative, on the implementation of Framework of Responsible State Behaviour in Cyberspace,[2] it is paramount to develop a more granular understanding of how AI is in practice changing capabilities and behaviours of both perpetrators (i.e., the attackers) and defenders during each step of malicious ICT activities.

This brief draws from multiple sources[3] to create a simplified model of these steps, and is intended to inform policymakers and diplomats engaged in international ICT security discussions. The proposed model, referred to as the **ICT Intrusion Path**, maps the various actions based on where they are taking place with respect to the targeted networks and examines them through the lens of AI's potential role in both malicious ICT acts and in the related defences.

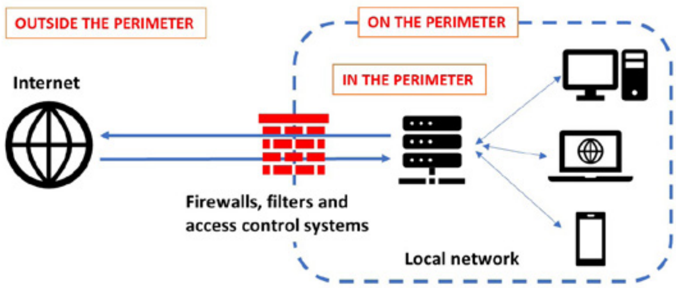
The research brief is structured as follows: after this introduction, **Section 2** provides a basic explanation of the concept of Network Perimeter, which is used as the main criterion to group the steps of the ICT Intrusion Path into three main categories: *outside the perimeter*, *on the perimeter* and *inside the perimeter*. **Section 3** and **Section 4** provide a first general introduction to the AI–ICT nexus from the perspectives of both perpetrators and defenders, respectively. **Section 5** provides initial reflections on how the AI–ICT nexus could be further explored in the context of current and future multilateral discussions on international ICT security. Finally, **Figure 1** and **Figure 2** illustrate the two sides of the ICT Intrusion Path, offensive and defensive, providing a more granular description of the impact of AI. Each figure also contains a dedicated glossary of key terms and definitions.

It should be acknowledged that, as AI technology is rapidly evolving, this brief is intended as a living document that should be updated as appropriate when new trends emerge. Also, it is important to note that in an effort to make this document useful for policymakers, some technical aspects are simplified.

2. Hold the Line: Introducing the concept of network perimeter

A useful framework to understand the unique challenges and opportunities that AI brings to both perpetrators and defenders of networks and systems is one that uses the perimeter of a network as a reference. Broadly speaking, 'network perimeter' refers to those systems that delimit a specific network from the broader Internet, mostly managing security of and access to internal networks (see Figure 1). This section introduces three key layers of analysis – **outside the perimeter**, **on the perimeter**, and **inside the perimeter** – providing for each both a simple definition and an overview of the most common subcomponents. Understanding these three layers will allow non-technical readers to better grasp the impact that AI may have on different individual actions as illustrated in the Infographic presented later in this brief.

Simplified overview of a network perimeter



Note on Cloud resources: *cloud resources are integral to modern networks,[4] often spanning both outside and on the perimeter. External services like public cloud platforms typically fall outside the perimeter, requiring robust policies for access and configuration security. In contrast, organization-controlled cloud infrastructure, such as hybrid clouds or cloud-hosted applications, functions on the perimeter, serving as critical access points and potential vulnerabilities. As extensions of the network, cloud resources demand tailored security strategies and shared accountability with providers. While this is an important topic, a detailed discussion of the security of cloud resources and the related impact of AI is outside of the scope of this paper.*

a. OUTSIDE the Perimeter:

This domain encompasses all the systems, networks, and data sources that exist beyond an organization's direct control. It includes public, external environments where perpetrators may gather intelligence on a target without actually interacting with its protected network. From a defender's perspective, this is where threat intelligence gathering takes place. Examples of relevant environments include:

- **public databases and repositories:** information on vulnerabilities, configurations, or even employee profiles can be gathered from online databases, software repositories, and code-sharing platforms;
- **social media and public websites:** publicly available information on employees, organizational structure, and technological dependencies can be obtained via social media profiles, press releases, and job postings;
- **dark web and cybercriminal forums:** dark web marketplaces can provide insight into new exploits, vulnerabilities, or prepackaged intrusion tools targeting specific systems or sectors; and
- **open-source intelligence (OSINT) sources:** from a defender's perspective, all of these external resources can be monitored to anticipate threats and manage vulnerabilities.

b. ON the Perimeter:

The perimeter represents the boundary between an organization's internal systems and the external world. This boundary is protected by layers of security meant to filter, monitor, and control access. Systems at the perimeter are usually configured to detect unauthorized access attempts and to protect the network from a wide array of external threats. Examples of systems found on the perimeter include:

- **firewalls and intrusion detection/prevention systems:**[5] these are stationed at the network's edge (as well as at other parts of the network) to monitor, filter, and potentially block malicious traffic;

- **email and content filters:**[6] these systems intercept potentially harmful content before it reaches internal networks, screening for phishing attempts, malware, or suspicious attachments; and
- **authentication and access control systems:**[7] these systems verify user identities and enforce access permissions and restrictions.

c. INSIDE the Perimeter:

Once within the perimeter, perpetrators have breached the internal network and can interact with critical systems, databases, and other sensitive assets. This domain is often characterized by a series of segmented and monitored internal networks that house sensitive data and operational systems. Examples of systems found in the perimeter include:

- **internal networks, data servers and file repositories:** these include servers hosting proprietary data, customer information, intellectual property, and other valuable assets, including classified or otherwise sensitive information, both military and civilian;
- **endpoint devices:** these are computers, mobile devices, and other equipment used by employees; and
- **network segmentation and monitoring systems:** these include systems that organizations implement to limit perpetrators' mobility and protect sensitive areas.

AI is a powerful tool that can be leveraged by both malicious actors and network defenders across all the three layers described above. The next two sections introduce the two use cases to provide context for the detailed explanation provided in the infographic. It is important to note that, particularly in relation to possible future developments, there is a degree of speculation as to what the impact of AI might be. What seems conceptually and theoretically possible, technological, financial, legal or other barriers may impact the actual transition from theory to practice.

3. INSIDE the AI–ICT Nexus: use cases for perpetrators

AI has become a formidable asset in malicious ICT activities, fundamentally changing how perpetrators approach, plan, and execute intrusions. That being said, given the current state of AI technologies, the utility of AI for malicious actors is not equally distributed across the three layers described in the previous section.

Traditionally, malicious ICT activities have required significant manual effort, from intelligence gathering to exploit creation. AI has transformed these processes, allowing perpetrators to operate with greater efficiency, adaptability, and stealth. Activities 'outside the perimeter' are where AI is currently providing the greatest advantages for perpetrators.

However, AI's impact is not limited to streamlining traditional techniques; it also opens the door to entirely new offensive approaches once the malicious actor reaches the network perimeter and seeks to penetrate the targeted network. This field of application of AI is rapidly growing, in particular thanks to the progress made in Generative AI.

Finally, AI's potential to adapt in real-time hints at a future where intrusion attempts can self-modify to counteract defensive responses, presenting a substantial escalation in cyber risk. However, this type of application at the time of writing remains the least developed and diffused. This is due to fundamentally two factors:

(i) AI models require substantial amounts of data to be trained, and by design malicious actors may not have access before an intrusion is attempted to enough specific data to train the model;

(ii) achieving this level of autonomy in malware, going beyond process automation (i.e. pre-programmed as rules based on Boolean logic like "if this is true, then do x; otherwise, do y"), would require deploying as part of the malicious payload an entire AI model, which, due to size and other parameters, would most likely be intercepted by various firewalls and intrusion detection systems (although this limitation might be overcome as smaller models are becoming more capable).

As a result, AI-driven offensive capabilities empower not only cyber operatives by enhancing their productivity but also lowers entry barriers for lower-skilled malicious actors by democratizing access to both knowledge and powerful and adaptable tools that could lead to potentially destabilizing effects on ICT security.

4. INSIDE the AI–ICT Nexus: use cases for defenders

As ICT threats grow in complexity and volume, AI emerges as a critical force multiplier for ICT defence. AI systems enhance defensive capabilities including threat detection and response, analysing vast data streams for unusual patterns, and reacting in real-time to potential intrusions. This capacity enables defenders to more effectively identify and mitigate threats before they cause significant harm, providing a strategic advantage in a rapidly shifting threat landscape.

In this context, while defenders can gain advantage in deploying AI across all layers of analysis, the relationship between utility and relative position with respect to the network perimeter is reversed when compared to perpetrators. In fact, the strongest use case today to deploy AI for ICT security is inside and on the network perimeter where vast volumes of data can be used to continuously train and improve defensive AI model(s).

As AI continues to be developed, these technologies will play an increasingly central role in helping organizations and governments defend their networks, bolstering cyber resilience in an era where digital infrastructure is foundational to global security.

5. The AI–ICT Nexus and the Framework of Responsible State Behaviour

Conducting a detailed assessment of the impact of AI on the Framework of Responsible State Behaviour is outside of the scope of this research brief and will be the subject of future research at UNIDIR. That said, it is possible to provide at least an initial overview of the main themes that could be subject to further discussions. For example, it would be useful for States and the multi-stakeholder community to:

a. bridge discussions on the application of international law and AI with discussions on international law and ICT security to identify any AI-specific challenges that may exist;

b. explore the impact of the AI–ICT nexus on norms in two directions – explore what specific challenges AI may bring to the implementation of existing norms and identify and leverage ways in which AI could be used to promote and facilitate such implementation;

c. explore how the combination of existing confidence-building measures, along with others potentially designed specifically for AI,[8] could support transparency and trust; and lastly;

d. consider AI as an important pillar of work for cyber capacity-building. This applies both to building capacity to mitigate AI as a new ICT security threat, but also to using AI to accelerate capacity-building, particular to increase cyber resilience, improve incident management and response, and mitigate the challenges arising from limited access to specialized skills.

UNIDIR INTRUSION PHASES

Figure 1

ICT INTRUSION PATH FOR PERPETRATORS

WITHOUT AI

OUTSIDE the perimeter

- The perpetrator manually sets/engages in various reconnaissance activities, including OSINT, social engineering, and scanning network infrastructure.
- The perpetrator must analyse and discern all the information gathered. All these activities can be time and resource consuming.
- The perpetrator prepares the resources for intrusion (manually, through automated tools,[10] or purchased), including exploits and malware code. Often, malware code is reused[11] with some modification.

ON the perimeter

- The perpetrator's ability to intrude into the victim's system relies on their skills in crafting delivery mechanisms (e.g., spear phishing emails) that bypass the victim's security capabilities (e.g., intrusion detection systems).
- Once inside the system, the perpetrator uses command-line scripts to remain there through various persistence techniques.[14]

INSIDE the perimeter

- Perpetrators manually set up 'command and control' within the compromised system.[18]
- Perpetrators work with command-line scripts and navigate through network directory service without being detected.
- Perpetrators can be supported by hacking tools to achieve their objectives.[19]

DESCRIPTION OF THE PHASES

OUTSIDE the perimeter[9]

The perpetrator gathers information on the victim that can be used to prepare the malicious ICT act (e.g., emails, network characteristics, vulnerabilities). Then, the perpetrator uses this information to prepare the resources to begin the intrusion, which consists of coupling malware with a vulnerability in the perimeter to create a payload to be used during the intrusion.

ON the perimeter[15]

The perpetrator attempts to penetrate the victim's system. The intrusion begins with the delivery of malware (e.g., through a phishing email or supply chain compromise) and the execution of exploits for the perimeter vulnerabilities identified in the previous phase. Once the perimeter is compromised, perpetrators must preserve access to the system.

INSIDE the perimeter[20]

The perpetrator penetrates the network and establishes command and control with the compromised system. The perpetrator can now carry out several actions according to their objective.

WITH AI

OUTSIDE the perimeter

- AI can support the analysis of vast amounts of information (e.g., dark web forums, social media), scanning network infrastructure, identifying patterns of behaviors of potential victims, and quickly producing comprehensive intelligence analyses.
- AI supports the perpetrator's skills and knowledge to prepare the resources for the intrusion (e.g., AI can support malware creation and/or customization, including 'polymorphic' malware, and ease the production of different exploits from the same vulnerability).[12]
- Possible developments: AI might be used to create new types of malware (e.g. DeepLocker).[13]

ON the perimeter

- AI has been used to craft and deceive sophisticated and customized products (websites, emails, and voice and video synthesis) that can be used to penetrate the network.
- AI can also support perpetrators in prioritizing and exploiting systems in large-scale intrusion attempts[16] and helping to evade detection.
- Possible developments: AI might increase malware code obfuscation for evasion and persistence.[17]

INSIDE the perimeter

- AI can support limited specific techniques (e.g., internal spear phishing or evasion of detection).[21]
- Possible developments: AI might be deployed inside the perimeter in the form of AI agents.

---

1 OEWG 2021–2025, Annual Progress Report, 12 July 2024, para. 22
2 See: https://nationalcybersurvey.cyberpolicyportal.org/background-to-un-discussions-on-responsible-state-behaviour
3 See, for example: https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html or https://attack.mitre.org
4 For more information on Cloud, please see: Federico Mantellassi, Giacomo Persi Paoli (2024). "Cloud Computing and International Security: Risks, Opportunities and Governance Challenges". UNIDIR, Geneva, Switzerland.
5 For further information on these, see, for example: https://www.infosecinstitute.com/resources/network-security-101/network-design-firewall-idsips
6 On these, see: https://learn.microsoft.com/en-us/exchange/antispam-and-antimalware/antispam-protection/content-filtering
7 See, for example, https://www.identity.com/the-role-of-authentication-and-authorization-in-access-control
8 Aloana Puscas (2022) "Confidence-Building Measures for Artificial Intelligence: A Framing Paper", UNIDIR, Geneva, Switzerland. Available at: https://unidir.org/publication/confidence-building-measures-for-artificial-intelligence-a-framing-paper/ vulnerability
9 This phase includes Reconnaissance and Resource Development tactics identified in the MITRE ATT&CK and Reconnaissance and Weaponization of the Cyber Kill Chain.
10 These tools often feature databases of exploits that perpetrators can search through to find ones that suit their victim's apparent vulnerabilities; see https://cset.georgetown.edu/publication/automating-cyber-attacks
11 Michal Tereszkowski-Kaminski, Santanu Kumar Dash, and Guillermo Suarez-Tangil. "A Study of Malicious Source Code Reuse Among GitHub, StackOverflow and Underground Forums." Computer Security ESORICS 2024, LNCS 14984, pp. 45–66, 2024.
12 Anonymous participants in the UNIDIR Workshop on ICT Intrusion Chain, 21 October 2024.
13 DeepLocker is a new type of malware that can target a specific victim and not others, as it is trained to recognize specific characteristics to become activated. Otherwise, it remains concealed.
14 For example, the perpetrator can install a backdoor to re-enable malware upon reboot or modify authentication mechanisms and processes to access user credentials.
15 This phase includes Initial Access, Execution, and Persistence tactics identified in the MITRE ATT&CK and Delivery, Exploitation, and Installation of the Cyber Kill Chain.
16 Jennifer Tang, Tiffany Saade, and Steve Kelly. "The Implications of Artificial Intelligence in Cybersecurity: Shifting the Offense-Defense Balance." The Institute for Security and Technology, October 2024.
17 Ibid.
18 It can involve choosing the communication protocol (HTTP, DNS, etc.) and its frequency, to avoid detection and to obfuscate traffic.
19 There are multiple hacking tools that support perpetrators in conducting several actions, such as lateral movement, compressing and encrypting files for safe transfer without detection, etc.
20 This phase includes Privilege Escalation, Avoid Detection, Credential Access, Discovery, Lateral Movement, Collection, Command & Control, Exfiltration, Impact – tactics identified in the MITRE ATT&CK and Command & Control and Actions on Objectives of the Cyber Kill Chain.
21 Maia Hamin and Stewart Scott. "Hacking with AI: the Use of Generative AI in Malicious Cyber Activity". Atlantic Council, February 2024.

Figure 2

ICT INTRUSION PATH FOR DEFENDERS

DESCRIPTION OF THE PHASES

OUTSIDE the perimeter

Defenders search for signs of malicious activities through different methods, including browsing vulnerability repositories, collecting server logs, and identifying reconnaissance behaviour.

ON the perimeter

The defenders must secure the perimeter and avoid intrusion. Defenders can rely on active or passive defensive measures. Both these measures are augmented by the information gathered outside the perimeter.

INSIDE the perimeter

Defenders must identify and block a perpetrator's command and control and rapidly detect and disrupt the perpetrator's actions to limit the impact of intrusion.

WITHOUT AI

OUTSIDE the perimeter

- Defenders manually set keywords to conduct OSINT regarding vulnerabilities and other sensitive information about their systems.
- Defenders subscribe to (and read) threat intelligence feeds.
- Malicious behavior is manually identified by analysing trends and patterns from gathered data.
- Some defenders' activities can be automated, but humans must still direct and supervise.
- All these activities can be time and resource consuming.

ON the perimeter

- Defenders rely on basic email filters and traditional antivirus solutions that often rely on known malware signatures.
- Defenders manually set up automated network traffic monitoring and filtering.[25]
- Defenders conduct regular vulnerability scanning and penetration testing on their systems. They must analyse and discern all the information gathered. All these activities can be time and/or resource consuming (e.g., as a service).
- Manually analysing ICT incidents to increase situational awareness.[26]
- User awareness training/testing.[27]

INSIDE the perimeter

- Defenders manually investigate the intrusion.[30]
- Responses can be delayed as defenders rely on basic monitoring tools and logs.
- Defenders must manually initiate containment steps, such as disconnecting affected devices, shutting down compromised accounts, or blocking malicious IP addresses.
- Defenders are often reactive, detecting intrusions only after they have progressed through the system.

WITH AI

OUTSIDE the perimeter

- AI is becoming valuable for defenders, particularly in quickly producing enhanced threat intelligence,[22] including learning from past incidents and predicting future intrusion vectors.[23]
- AI can detect payload weaponization.[24]
- Possible developments: AI might be developed to shift defensive strategies towards preemptively stopping threats before they emerge.

ON the perimeter

- AI can enable better perimeter protection, including through:
  - enhancing email filters and antivirus solutions[28]
  - hardening authentication systems[29]
  - supporting (continuous) penetration testing and fixing vulnerabilities
  - setting up honeypots that mimic real network behavior
  - supporting scan tools that more precisely identify reconnaissance patterns and suspicious behaviour.
- Decentralized AI cybersecurity solutions are emerging.

INSIDE the perimeter

- AI rapidly collects, processes, and summarizes useful intelligence on unusual/suspicious activities within the network.
- AI-enabled security helps to prioritize tasks and guides defenders in taking actions to contain, prevent, and mitigate intrusions.[31]
- AI tools can compile detailed reports on ICT intrusion necessary for reporting ICT incidents.
- Possible developments: AI developments might enable 'self-healing' networks and systems that preserve normal operations in the event of malicious intrusions.

22 For example, at Microsoft, one team of analysts takes one week to identify and process 50 articles; with AI, the team can now generate concise reports from these articles in minutes; see Microsoft. "Microsoft Digital Defense Report 2024". Microsoft, 2024.

23 "AI can be considered as a sense-making tool" participants at the UNIDIR closed-door workshop on ICT Intrusion Chain, 21 October 2024; Jennifer Tang, Tiffany Saade, and Steve Kelly. "The Implications of Artificial Intelligence in Cybersecurity: Shifting the Offense-Defense Balance." The Institute for Security and Technology, October 2024.

24 There are tools that work by detecting suspicious patterns and behaviours during code development and modification stages and can inform defenders on potential exploitation behaviour.

25 These tools (e.g., firewall, honeypot, proxy servers, etc.) allow for automated blocking of packets based on pre-configured rules.

26 Traditional methods include manually analysing the malware that was sent to the infected system and running it in a closed simulated network (sandbox) to understand its actions. These analyses are likely time-consuming, even for the most highly qualified ICT analysts, and may have limited network simulation capabilities; see Napoleon C. Paxton et al. "Utilizing Network Science and Honeynets for Software Induced Cyber Incident Analysis". 48th Hawaii International Conference on System Sciences. 2015.

27 Lockheed Martin. "Gaining the Advantage. Applying Cyber Kill Chain Methodology to Network Defence". Lockheed Martin 2015.

28 These tools sentiment analysis and semantic parsing for compliance. Moreover, they rely on AI for proactive security, especially in environments where real-time, automated response is critical for robust endpoint protection.

29 For example, AI enables continuous authentication by monitoring user behaviours throughout their session; see Madison Evans. "How AI is Revolutionizing User Authentication Systems". Eartho, https://www.eartho.io/blog/how-ai-is-revolutionizing-user-authentication-systems

30 Traditional security often relies on rules-based systems and alerts. Defenders must analyse each alert to identify malicious activity. This process can be time-consuming, and in case of multiple, complex, or subtle intrusions, alerts may be missed.

31 Anonymous participants in the UNIDIR Workshop on ICT Intrusion Chain, 21 October 2024; Microsoft. "Microsoft Digital Defense Report 2024". Microsoft, 2024.

Glossary

Active Defense	proactively detects and diverts intrusion attempts (e.g., MITRE Engage framework)
AI Agent	an AI-based system or program that autonomously performs tasks, for example related to detecting, analysing, preventing, and responding to cybersecurity threats
Code obfuscation	the process of modifying code to make it harder to be analysed and reverse engineered
Command-line	text-based interface that allows users to interact with a computer by typing commands
Decentralized cybersecurity (with AI)	solutions that enable real-time, secure, and anonymous sharing of threat data among users to adapt collectively to evolving threats
Directory Service	a database containing information about users, devices, and resources
Exploit	a program or piece of code designed to advantage of a security flaw or vulnerabilities
Honeypot	is a decoy system or network designed to attract, detect, and study malicious ICT activities by mimicking a real target
Malware signature	a unique pattern or code snippet used to identify specific malware
Passive Defense	focuses on denying access on the perimeter by detecting and blocking suspicious activities (e.g., using firewalls, scanning ports, restricting administrator privileges)
Payload	a component of malware that executes malicious actions, such as data theft, encryption, or system disruption
Phishing	a ‘social engineering’ technique where deceptive emails, messages, or websites are used to trick individuals into revealing sensitive information, such as passwords or financial details
Polymorphic malware	malicious software that adapts its code or behaviour to evade detection by traditional security tools like antivirus programs
Social Engineering	the manipulation of people through deception, persuasion, or psychological tactics to gain unauthorized access to sensitive information
Spear phishing	a targeted form of phishing aimed at specific individuals or organizations, often using personalized information to appear more convincing and increase the likelihood of success
Vector	the method used by the perpetrator to penetrate the system or network of a victim (e.g., phishing)
Vulnerability	a flaw in an ICT system that can be used by a perpetrator to achieve their objectives
Weaponization	the act of developing and combining malware and an exploit in a payload to be delivered to a victim’s system
Weblogs	files that record activity or events on a web server, including who requests to access it, when, and from where

@unidir
/unidir
/un_disarmresearch
/unidirgeneva
/unidir

UNIDIR

Palais de Nations
1211 Geneva
Switzerland

WWW.UNIDIR.ORG

Authors: Giacomo Persi Paoli, Samuele Dominioni.

Support from UNIDIR core funders provides the foundation for all of the Institute’s activities. This study was produced by UNIDIR’s Security and Technology Programme with the support of Czech Republic, France, Germany, Italy, the Netherlands, Norway, Republic of Korea, Switzerland, and Microsoft.

The authors wish to thank the following individuals for their invaluable advice and assistance on this report: Ambra Demontis, Gaurav Keerthi, Ivan Zelinka, Kazuo Noguchi, Josiah Hagen, Michael Karimian, Myles Foley, Peter Slattery.

Design: Kathleen Morf, www.kathleenmorf.ch
Photos: © Adobe Stock and ChatGPT/AI Generated.

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.

Note