WASHINGTON INTERNATIONAL LAW JOURNAL

The Weaponization of Artificial Intelligence in North Korean Military Espionage

By: **Bhavya Johari** | October 23, 2025

North Korea's employment of Artificial Intelligence (AI) to forge military identification documents represents a critical inflexion point in state-sponsored cyber warfare. The September 14, 2025 disclosure that the North-Korean state-sponsored Kimsuky Advanced Persistent Threat (APT) group employed ChatGPT to create deepfake South Korean military credentials for espionage operations signals not mere tactical innovation, but a fundamental transformation in how smaller state actors can project sophisticated intelligence capabilities. This development urgently requires reconsidering existing legal frameworks governing artificial intelligence applications and international cybersecurity protocols.

The Technical Evolution of State Deception

The Kimsuky group's operation demonstrates how commercial AI platforms can be weaponised for state-level intelligence gathering without requiring indigenous technological development. Cybersecurity researchers identified that the military ID images were the deepfakes with a 98% probability, which were then embedded within phishing emails designed to impersonate legitimate South Korean defense communications. The attackers circumvented ChatGPT's built-in restrictions against generating government identification by reframing their requests as sample designs for legitimate purposes rather than direct reproductions.

This tactical evolution transcends traditional social engineering by manufacturing visual evidence of authority rather than merely impersonating officials through text or voice. The sophistication of these AI-generated credentials demonstrates how readily available tools can produce materials indistinguishable from authentic government documents, fundamentally undermining trust mechanisms that underpin institutional communications. When state actors can generate convincing visual authentication materials using commercial platforms, the assumption that official documentation provides reliable identity verification collapses.

Legal Frameworks Inadequate for AI-Enhanced Espionage

Current international legal frameworks governing cyber warfare and artificial intelligence misuse are inadequate to address AI-enhanced state espionage. The European Union's AI Act, which entered force in August 2024, represents the most comprehensive regulatory attempt yet fails to address state-sponsored misuse of AI tools developed by third parties. While the Act mandates transparency for AI-generated content and establishes disclosure requirements, it provides no mechanisms for preventing malicious state actors from circumventing these requirements or holding them accountable for misusing legitimate AI platforms.

Traditional espionage law struggles to categorize AI-generated identity fraud within existing offence definitions. When North Korean actors use American-developed AI tools to forge South Korean military documents for intelligence gathering, questions of jurisdiction, attribution, and proportional response become extraordinarily complex. The incident reveals how AI democratizes sophisticated deception capabilities, enabling smaller state actors to deploy previously resource-intensive operations with minimal technical infrastructure.

The absence of clear legal precedent for prosecuting AI-enhanced state espionage creates dangerous operational ambiguity. International law operates within frameworks developed before AI-generated content became accessible to state actors, leaving substantial gaps in regulatory coverage and enforcement mechanisms.

South Korean, right, and North Korean army soldiers stand guard at the border village of Panmunjom in the demilitarized zone (DMZ) between the two Koreas. Ahn Young-joon/AP

Attribution Challenges in Synthetic Media Operations

The employment of deepfake technology fundamentally complicates attribution mechanisms essential for international law enforcement and diplomatic response. Traditional cyber-attack attribution relies on technical indicators, behavioral patterns, and digital forensics. However, AI-generated content introduces new layers of plausible deniability that existing attribution methodologies cannot adequately address.

The Tallinn Manual 2.0's principles for applying international law to cyberspace assume states can be reliably identified as responsible for cyber operations. AI-generated content threatens this foundational assumption by making definitive attribution more challenging. When state actors can generate synthetic evidence indistinguishable from authentic materials, meeting the evidentiary standards required for international legal proceedings become nearly impossible

Current international law demands that states taking countermeasures correctly attribute internationally wrongful acts, imposing strict liability standards. This creates an impossible burden when dealing with AI-enhanced operations, where recursive doubt can be introduced about the authenticity of any digital evidence. If detected, operators can claim their synthetic materials were themselves targets of deepfake manipulation, creating layers of uncertainty that undermine legal accountability mechanisms.

Strategic Implications of Democratized Deception

North Korea's integration of commercial AI tools into state espionage operations represents a paradigm shift in how technologically smaller powers can project cyber capabilities traditionally reserved for major nations. The democratization of technology that can be used for sophisticated deception poses acute challenges for targeted nations' defensive planning. Military and intelligence agencies must assume that any visual or audio authentication could be AI-generated, fundamentally altering operational security protocols.

The psychological impact extends beyond immediate security concerns and erodes trust in digital communications systems essential for modern military cooperation and intelligence sharing. This uncertainty creates operational paralysis where institutions cannot reliably verify the authenticity of critical communications, potentially disrupting time-sensitive security operations.

The incident also highlights how private AI companies inadvertently become enablers of state-sponsored espionage. OpenAI's ChatGPT, designed for legitimate commercial applications, becomes a tool for manufacturing fraudulent government documents when accessed by hostile state actors. This raises urgent questions about AI developers' responsibilities to prevent platform misuse and the feasibility of technical safeguards against sophisticated state-level abuse.

Establishing Multilateral Attribution Bodies and Binding AI Platform Safeguards

Addressing AI-enhanced state espionage requires immediate and concrete policy interventions that must be implemented to prevent the normalization of synthetic media operations in statecraft.

The international community must establish an independent multilateral attribution mechanism for cyber operations, as recommended by the UN Secretary-General. This body must develop evidentiary standards for attribution challenges posed by AI-driven attacks, moving beyond traditional technical indicators to account for synthetic media operations where conventional forensic methods prove inadequate.

AI platform developers must be subject to binding technical safeguards that prevent state-level misuse. This includes mandatory implementation of robust verification protocols following never trust, always verify principles and adopting the National Security Agency's Artificial Intelligence Security Centre (AISC) data integrity guidelines specifically designed to detect and prevent malicious use patterns characteristic of state intelligence operations. Corporate responsibility cannot remain voluntary when commercial platforms become instruments of international espionage.

Finally, nations must develop international cooperation protocols for deepfake detection and response, enabling rapid cross-border investigation of synthetic media cyber operations. These protocols must include information-sharing agreements, joint technical analysis capabilities, and coordinated diplomatic responses that hold state actors accountable for AI-enhanced espionage while respecting sovereignty concerns.

The window for preventive action closes rapidly as AI capabilities proliferate and technical barriers diminish. Without these specific mechanisms implemented immediately, AI-enhanced espionage becomes normalized statecraft, permanently undermining the digital authentication systems essential for international security and diplomatic relations.

Bhavya Johari is a Lecturer at Jindal Global Law School, O.P. Jindal Global University, India. He earned his undergraduate law degree from NALSAR University of Law, Hyderabad, India, and holds an LL.M. from Melbourne Law School, University of Melbourne, Australia.

Tags: Artificial Intelligence, ChatGPT, Cyber Espionage, DPRK, Espionage, Malware, North Korea, South Korea, Spying

Categories: International Law, Technology & Cyberlaw

Box 353020
Seattle, WA 98195-3020

Fax: (206) 685-4457
Email: winlj@uw.edu