PARIS PEACE FORUM

INTAiC

Bridging the Evidence Gap on AI Misuse in Cyberspace

*Discerning **signal** from noise on AI-enabled cyber threats*

INAUGURAL WHITE PAPER

JULY 2026

Foreword

This White Paper inaugurates the first pillar of the Paris Peace Forum's Integrated Network for Trusted AI in Cyberspace (INTAiC), a global, multistakeholder effort to address gaps in our shared, evidence-based understanding of how advanced artificial intelligence (AI) is being misused by adversaries in cyberspace. It builds on a sustained programme of consultations and collective reflection conducted by the Paris Peace Forum with frontier AI labs, the cybersecurity industry, international organizations, civil society and academia since 2024.

This endeavor does not seek to duplicate existing initiatives. It builds on the proven track record of public–private cyber cooperation, industry partnerships for securing AI, disclosure efforts of frontier model developers, and academia-led efforts to provide the scientific foundation for an international consensus on AI security. It aims to act where no single actor can act alone: at the intersection of model-side telemetry, downstream cyber observation, and international policy coordination.

As the workstream's co-leads, the Paris Peace Forum and the General Purpose AI Policy Lab are grateful to the partner organizations whose expertise has informed this document. Some of their contributions are reflected in the quotations and institutional insets embedded throughout the text. **INTAiC remains open to all stakeholders who share the conviction that a free, open, secure and non-fragmented cyberspace cannot be sustained without a credible response to the AI-enabled threat landscape**.

Tom David, CEO, GPAI Policy Lab

Pablo Rice, Head of Programme, Paris Peace Forum

Executive Summary

Advanced artificial intelligence has crossed a threshold in cyberspace. What was, only eighteen months ago, an emerging research curiosity — the use of large language models by adversaries to assist reconnaissance, draft phishing lures or generate fragments of malicious code — has matured into an industrial-scale capability, with myriad benefits for human society. At the same time, criminal and state-affiliated actors attempt to use proprietary and open-weight models for end-to-end attack lifecycle support, including the autonomous discovery of zero-day vulnerabilities, the dynamic generation of obfuscated malware, agentic orchestration of intrusion campaigns, and supply-chain compromise of AI components themselves. A parallel underground economy of brokered AI capabilities is consolidating in the dark-market spheres, with specialized vendors lowering the cost and skill floor of conducting offensive operations.

The defensive side is closing in. New AI-native security systems and structured disclosure programmes demonstrate that AI can also operate as a force multiplier for defenders. Yet the asymmetry remains structural: attackers benefit from rapid experimentation without legal or reputational consequences, while defenders are constrained by patch cycles, regulatory reporting requirements and the fragmented visibility that each constituency has into the AI-cyber-attack chain.

This White Paper argues that the most consequential governance gap is not a capability gap but an evidence gap on actual misuse. For most purposes, the use of AI by attackers can now be treated as a baseline assumption. The questions of greater consequence for policymakers are how much capability it adds and what effects it produces, rather than how frequently it is used. Over the past three years, most efforts have concentrated on the upstream assessment of frontier AI cyber capabilities, through evaluation frameworks, capability thresholds and pre-deployment safety testing. The downstream face of that work — the structured, transnational reporting of how, and with what impacts, those capabilities are actually being misused "in the wild" — has lagged behind. Bridging that asymmetry is the operational core of this initiative.

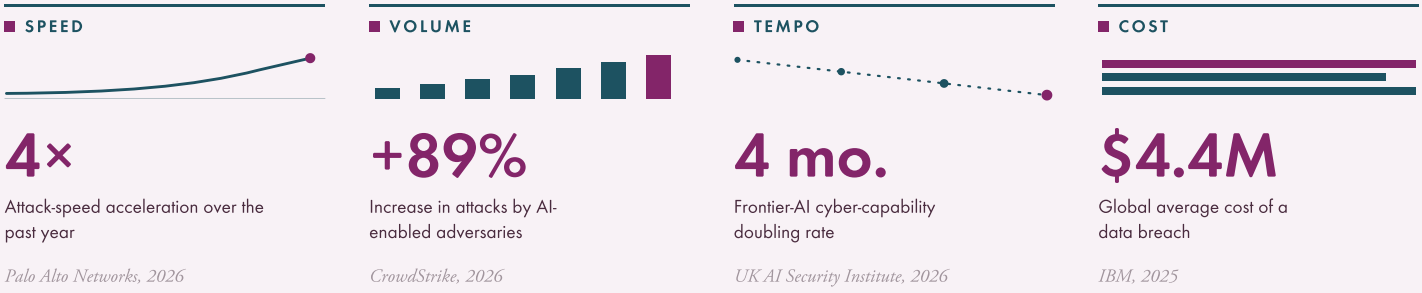
INTAiC proposes a transnational collaborative framework that rests on three pillars: a shared analytical effort to map AI misuse in real world settings across complementary observation points; a briefing programme designed to equip governments, AI safety and security institutes, and international organizations with a common reading of the phenomenon; and a coalition-extension effort to bring jurisdictions and sectors so far under-represented in existing arrangements into the documentation work. The initiative will deliver its first mapping report and a series of briefings by the time of the next Paris Peace Forum's annual edition in November 2026.

Part One

The Accelerating AI-Cyber Threat Landscape

1.1 From experimentation to industrial-scale misuse

The trajectory of AI misuse in cyberspace has moved decisively from experimentation to industrialization. The shift can be approached along four converging dimensions (speed, volume, tempo and cost) and the early 2026 reporting cycle has produced consistent quantitative anchors for each of them.



On speed, the time elapsed between initial access and data exfiltration in the fastest observed intrusions has collapsed to a median of seventy-two minutes, an approximate fourfold acceleration relative to the previous year, in a corpus of more than seven hundred and fifty high-stakes incidents reviewed by leading incident-response practitioners. Time-to-exploit on critical vulnerabilities is collapsing in parallel, with industrialized tooling sold on dark-web marketplaces and Telegram channels reducing reconnaissance-to-exploitation windows that once spanned weeks to single-digit hours [1].

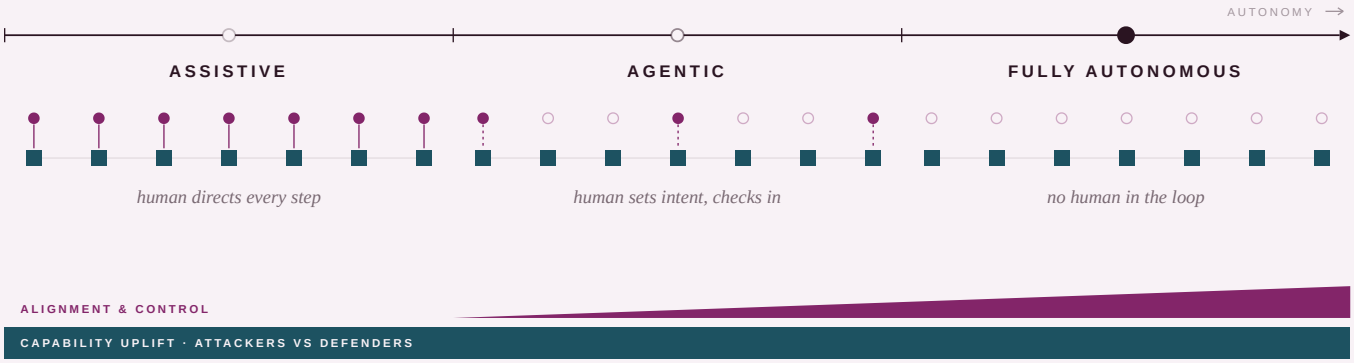
On volume, AI-enabled adversary operations rose by approximately eighty-nine per cent year on year in 2025, according to consolidated industry telemetry covering more than two hundred and fifty tracked threat actors [2]. The increase is driven both by long-tracked state-affiliated clusters that have absorbed generative AI into their existing tradecraft and by the diffusion of AI-augmented commodity tooling across the cybercriminal ecosystem. Generative AI is now embedded across reconnaissance, social engineering, scripting, credential theft and operational execution, to the point where the relevant question is no longer where AI is used in the attack chain but where it is not.

On tempo, the underlying capability is itself accelerating. Independent evaluations conducted by national AI security institutes indicate that the autonomous cyber-task horizon of frontier models, namely the length of cyber tasks they can reliably complete without human intervention, was doubling approximately every four months at the start of 2026, against an estimate closer to eight months only three months earlier [3]. The most recent frontier models evaluated in the spring of 2026 have significantly outpaced even that revised trend, although it remains

too early to determine whether this constitutes a further structural acceleration or an isolated step change. Either way, this second-derivative dynamic compresses the time available to policymakers, defenders and operators to internalize each new capability plateau before the next one arrives, and demonstrates a distinct imperative to leverage AI capabilities in the service of network defense. At the current pace of progress, most benchmarks currently used to evaluate frontier models' cyber capabilities might be saturated before 2030. Benchmark saturation in this domain is likely to bring the field closer, in the years ahead, to a stage where AI systems can automatically identify and exploit vulnerabilities in systems that have not been secured in time [4].

On cost, the macro-economic footprint of these dynamics remains difficult to attribute with precision, and most published estimates do not yet isolate the AI-related component of incident costs from the broader cybersecurity baseline. The three preceding observations should nonetheless be read against the order of magnitude of that baseline. The global average cost of a data breach stood at approximately USD 4.4 million in 2025, with the United States average exceeding USD 10 million [5]. Even modest amplifications of the speed, volume and tempo curves described above would translate, at this baseline, into materially higher aggregate losses, and quantifying that translation is precisely one of the empirical questions the Initiative seeks to address.

These four dimensions describe the intensity of AI misuse. **A separate consideration is its degree of autonomy, for which three modes can be distinguished** [6]. The first and best documented is assistive use, in which models support reconnaissance, phishing, scripting and social engineering while remaining under human direction. The second is agentic use, in which systems chain tasks and operate tools with partial autonomy. The third is fully autonomous operation conducted end to end without meaningful human involvement, which remains uncommon in operational settings owing to constraints of access, reliability and control. This threshold has now been crossed under controlled conditions. In a recent evaluation, a frontier model became the first to complete a full thirty-two-step corporate-network attack chain end to end without human intervention, succeeding in three of ten attempts [7]. The test environment included neither live defenders nor real-time response; the result therefore establishes autonomous compromise of weakly-defended systems, not of hardened enterprise networks. Such findings indicate the trajectory of capability, while assistive and early-agentic use continues to dominate observed operations. The two should not be conflated.



Each step along this progression reshapes the governance problem. While assistive use raises the question of how much AI lifts the capabilities of attackers, agentic use adds a further concern. As models move from executing human intent to choosing and sequencing their own actions, the question is no longer only how much capability AI adds, but whether the system's behaviour itself can be reliably controlled over extended operations. Responsibility for the resulting harm becomes harder to assign, and external scaffolding can extend an agent's reach beyond what model-level evaluation captures. This shift reflects a structural feature of how frontier models are built: a gap between the behaviour developers intend and what models ultimately exhibit in deployment. As systems become more capable and more autonomous, that gap widens.

Unlike misuse, which is inherently asymmetric between attackers and defenders, the control problem cuts across that distinction. Whether deployed for offensive or defensive purposes, an autonomous system may deviate from its intended behaviour in ways that neither its developer nor its operator anticipate. Here, questions of capability increasingly meet the broader agenda on AI alignment and control.

AN INDUSTRY PERSPECTIVE – **GOOGLE THREAT INTELLIGENCE GROUP**

"Threat actors are already leveraging AI for a meaningful boost to the speed, scale, and sophistication of their operations. We will have to overhaul the way we do defense to meet this challenge."

John Hultquist – Chief Analyst, Google Threat Intelligence Group

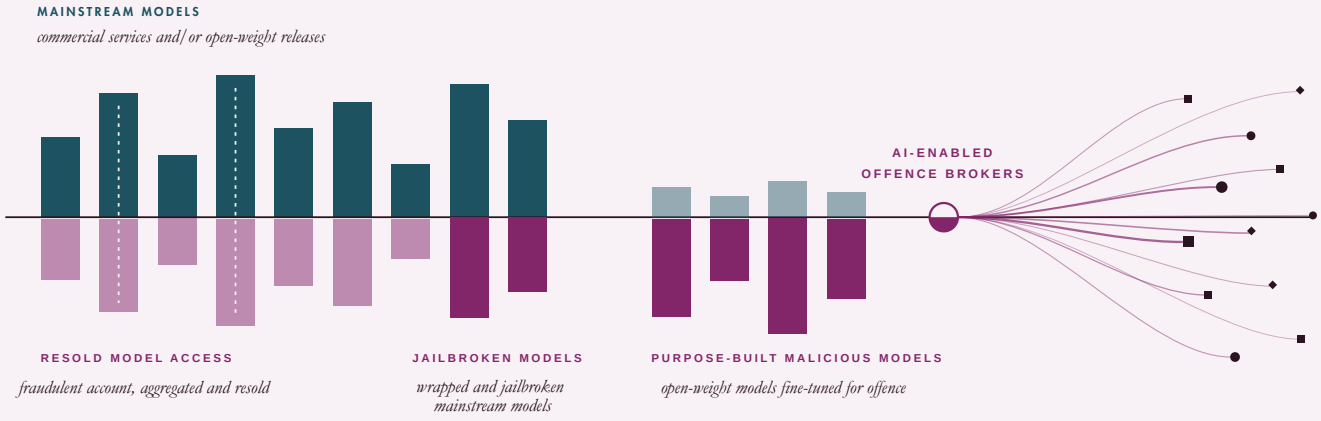
Beyond these aggregate trends, two structural shifts deserve particular attention:

- **The lifecycle of software vulnerabilities is shifting**. Models are increasingly used to find new flaws automatically and to turn security patches into working exploits, compressing the interval between disclosure and exploitation from weeks to days or hours. The pressure moves from vulnerability discovery to remediation triage – the order in which defenders decide which flaws to patch first. With tens of thousands disclosed each year, the task is no longer to find vulnerabilities but to rank them by how likely they are to be exploited. Public sources that track which flaws are under active attack are a key input, and one that broader trend analysis will also need.

- **Malicious activity is migrating toward open-weight models**, where defenders progressively lose visibility into adversaries' queries and experimentation, reducing the share of AI-enabled offensive activity that remains observable through provider-side telemetry. The inset that follows traces how both dynamics are playing out across state-affiliated and criminal threat clusters.

1.2 A new brokering economy: the underground market for AI-enabled offence

Alongside the visible evolution of frontier capabilities, a parallel underground economy of AI-enabled offence has emerged and consolidated since 2023. It is structured around three converging brokering patterns whose combined effect is to lower the cost and skill floor required to conduct offensive cyber operations.



First, the clandestine intermediation of access to legitimate frontier models. A grey-market layer of intermediaries has emerged, reselling anonymised, premium-tier access to mainstream commercial AI systems, typically by aggregating pools of fraudulently created accounts. Adversaries thereby obtain industrial-scale use of state-of-the-art capabilities while bypassing the safety guardrails, abuse monitoring and accountability mechanisms that frontier developers are otherwise expected to maintain. The result is a hybrid zone of partial visibility, neither fully inside nor fully outside the monitored frontier ecosystem.

Second, the brokering of jailbroken or purpose-built malicious models. From the original WormGPT, which emerged in mid-2023 as one of the first commercialised malicious LLMs, to successor and copycat services such as the resurgent WormGPT 4 brand and the free, lightweight KawaiiGPT, this category has matured from short-lived jailbroken chatbots into commercialized, subscription-based offerings advertised on underground forums and distributed through private messaging channels [8]. Industry tracking estimates that mentions of malicious AI tools on underground monitoring platforms grew by approximately two hundred per cent year on year between 2023 and 2024, with continued expansion observed into 2025 [9].

Third, the convergence between AI-enabled offerings and the established Initial Access Brokers (IABs) ecosystem. IABs are specialized actors that obtain unauthorized access to organizations' networks and resell it to ransomware, espionage or state-aligned operators. They have become a structural feature of the cybercriminal value chain, with documented surges in sectors such as healthcare, education, transport and public administration where access to critical systems is increasingly sold to the highest bidder [10].

The strategic implication of this brokering economy is twofold. It could accelerate the diffusion of capabilities that, until recently, required nation-state resources, and it does so within commercially structured channels that are responsive to demand and price signals. Furthermore, it crystallizes a new category of actors, brokers, whose position in the attack chain sits between the model developer and the intended user. This additional layer is one of the defining features of the current AI-enabled offensive economy.

AN INTERNATIONAL POLICING PERSPECTIVE — **INTERPOL (1/2)**

From an international policing perspective, the misuse of AI in cyberspace is best understood as an accelerant of existing criminal markets rather than a separate category of threat. Across INTERPOL's member countries, the most visible effects are emerging where cybercrime is already most scalable and profitable: online scams, phishing, business email compromise, identity theft, sextortion, ransomware, infostealer malware and other forms of cybercrime-as-a-service. In these areas, AI does not need to create wholly new offences to have a significant impact. Its immediate value to criminals lies in increasing the speed, volume, credibility and personalization of attacks, while lowering the technical and linguistic barriers that previously limited some offenders. Recent INTERPOL regional cyber threat assessments reflect this pattern:

ASIA & SOUTH PACIFIC – 2026

The 2026 INTERPOL Asia and South Pacific Cyber Threat Assessment Report highlights the rise of AI-enabled deepfake scams, industrial-scale fraud operations, ransomware and infostealer malware as key concerns. It also points to a surge in cyber-enabled criminal operations across Southeast Asia, including large-scale scam centers operated by transnational organized crime groups. These operations show that cybercrime in the region has a significant human impact, including exploitation, coercion and victimization that go far beyond financial losses.

AFRICA — 2025

INTERPOL's 2025 Cyberthreat Assessment (2) also shows how AI-enabled cybercrime is shaping the continent's existing cyber threat landscape. Malware, especially ransomware, online scams, phishing and business email compromise remain dominant threats, while AI-driven fraud, online image-based sexual abuse, digital sex crimes and cybercrime-as-a-service are emerging concerns. The report highlights digital sextortion as a particularly prominent issue — a clear example of how AI can intensify victim harm, especially where stigma, underreporting and limited investigative capacity already make these crimes difficult to address.

For INTERPOL, the challenge is not simply to identify whether AI was used, but to understand how it changed the nature, scale and impact of the crime. In many cases, AI may be one enabling factor within a broader criminal scheme, rather than the defining feature of the case itself. This makes it difficult to compare trends across jurisdictions, assess the added criminal capability created by AI, and determine what forms of cooperation, disruption or victim protection are most effective.

AN INTERNATIONAL POLICING PERSPECTIVE – **INTERPOL (2/2)**

A stronger international picture will therefore depend on common terminology, structured reporting and trusted channels through which operational signals can be shared and analyzed, while protecting ongoing investigations, personal data, sensitive techniques and national procedures. In this regard, INTERPOL's role is to help connect frontline law enforcement observations with the broader policy, operational and cybersecurity ecosystem.

« AI is a double-edged sword for law enforcement. It can help us be faster and more effective, but criminals are also using it to make their activities more convincing and scalable. Our priority is to pay attention to what member countries are seeing on the ground, turn that information into useful intelligence, and work with partners to disrupt threats and protect victims. »

Neal Jetton – Cybercrime Director, INTERPOL

1.3 Potential implications for national and international security

As offensive AI capabilities continue to accelerate and the underground economy that supports their diffusion matures, significant security consequences could emerge at two interlocking levels.

At the national scale, the response capacity of all but the best-resourced public and private defenders could be progressively outmatched. The response playbooks of national CERTs, sectoral regulators and critical-infrastructure operators would tend to be overwhelmed, leaving public administrations and supply chains exposed to a pace of intrusion they were not originally designed to absorb. The same pressure bears hardest on those least equipped to withstand it – civil-society organisations, small enterprises, schools, hospitals and local administrations – whose defences are thinnest precisely where attacks are becoming cheapest to mount. As these turn into the path of least resistance, the resilience of the wider system comes to rest on its least defended parts.

At the international scale, AI may be reshaping the global balance of cyber power through two opposing dynamics. On the one hand, capabilities that recently required the resources of a few cyber-mature powers may come within reach of second-tier actors with limited offensive traditions, lowering the threshold of cyber-enabled coercion in inter-state competition. On the other hand, a defensive polarization could intensify, with the risk of an even wider gap than before between societies able to deploy AI-backed defenses and those that cannot.

Cyber deterrence itself would come under unprecedented strain. As decision windows compress, the risk of premature reactions in a crisis increases. By the same token, the confidence-building measures developed over

the past decade to limit such escalations would be put to acute test. Addressing this set of pressures would likely require a response that runs from organisation-wide security postures all the way up to high-level measures within multilateral frameworks.

Whether these dynamics ultimately favour attackers or defenders remains unsettled. Attackers face fewer constraints in the short term, while some practitioners consider that the larger long-term advantages will accrue to defenders. This White Paper treats the question as one to monitor rather than to resolve.

AN INDUSTRY PERSPECTIVE – **ORANGE CYBERDEFENSE**

"AI will doubtlessly accelerate the discovery and exploitation of vulnerabilities yet again raising the bar for defenders, especially if they are isolated from one another. But despite this rapidly evolving landscape where changes are happening almost daily, new attack techniques do not necessarily equal new vulnerabilities or necessitate new strategies to be implemented by threat actors. Therefore, having a firm yet adaptable strategy is essential.

Our response to this escalation must be carefully considered, not reflexive. Organizations seeking to adopt AI tools must adjust first their risk models, keeping in mind the security and integrity of their models, but also recognizing that data itself is a critical element of sovereignty and geopolitical influence. It is then essential to consider the data feeding these systems, the controls around access, the reliability of outputs, and the governance of AI behaviours. Since it is almost certain that we are going to face an increasing amount of diversified and sophisticated cyberthreats, it is essential that security practitioners, be it in cyber, AI and government, share their observations and best practices to reinforce our collective immunity."

Olivier Bonnet de Paillerets – Executive Vice-President, Strategy and Anticipation, Orange Cyberdefense

Part Two

A Remaining Evidence Gap: The Next Imperative for Governance

2.1 Capturing the full spectrum beyond capability assessments

Any international consensus on AI-cyber security will require an empirical foundation that does not yet exist. The International AI Safety Report 2026 identifies this evidence gap as one of the central policy challenges at the AI-cyber intersection. The effectiveness of AI misuse, the additional capability it provides to attackers and the consequences it produces all remain difficult to quantify. These dimensions are more relevant to policy than the overall prevalence of AI use by adversaries. In their absence, policymakers must choose between acting on incomplete information and waiting for data that may take years to consolidate [11].

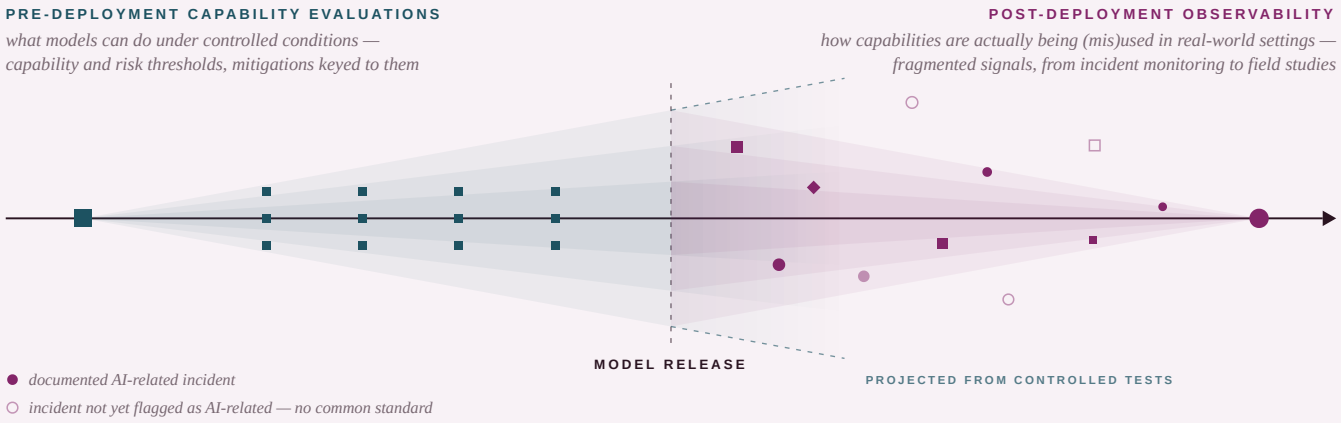
This gap is widening within a context of unusually rapid technical evolution. AI-cyber capabilities now move on time-scales of weeks, while the international processes on these issues operate on multi-year cycles. The resulting mismatch is structurally different from the gap that has long existed between ICT governance and the technology it sought to address, and it creates a real risk of normative vacuum.

Governance efforts to address this challenge have, in recent years, concentrated on the assessment of what advanced models can do under controlled conditions. Frontier developers and national AI Safety and Security Institutes have converged on a common methodological core organized around cyber threat modelling, pre-deployment capability evaluations, capability and risk thresholds, and mitigation measures keyed to those thresholds [12]. This work has substantially improved the international understanding of what frontier systems are technically capable of in the cyber domain.

By design, however, controlled evaluation can capture only part of the picture. Pre-deployment evaluations are predominantly conducted in controlled testing environments that cannot account for real-world dynamics, and the most advanced AI systems are increasingly embedded in larger workflows that current benchmarks do not measure. These methodological challenges are well recognized within the evaluation community itself [13], and they make a complementary effort all the more important. Closing this gap will indeed require a parallel programme of structured monitoring of how these capabilities are actually being (mis)used in real-world settings.

Understood in these terms, post-deployment observability is itself a governance priority. It spans both the emerging efforts to evaluate models after release and the structured documentation of how they are misused across the attack chain. The difficulty is no longer in understanding a single intrusion, but in observing such

cases clearly and continuously, then bringing them together. No agreed standard yet exists for flagging a cyber incident as AI-related, and no common infrastructure for tracking such cases across operations and over time. Building that capacity represents the next frontier, best approached as a single, connected challenge rather than left to each defender and each jurisdiction in isolation.



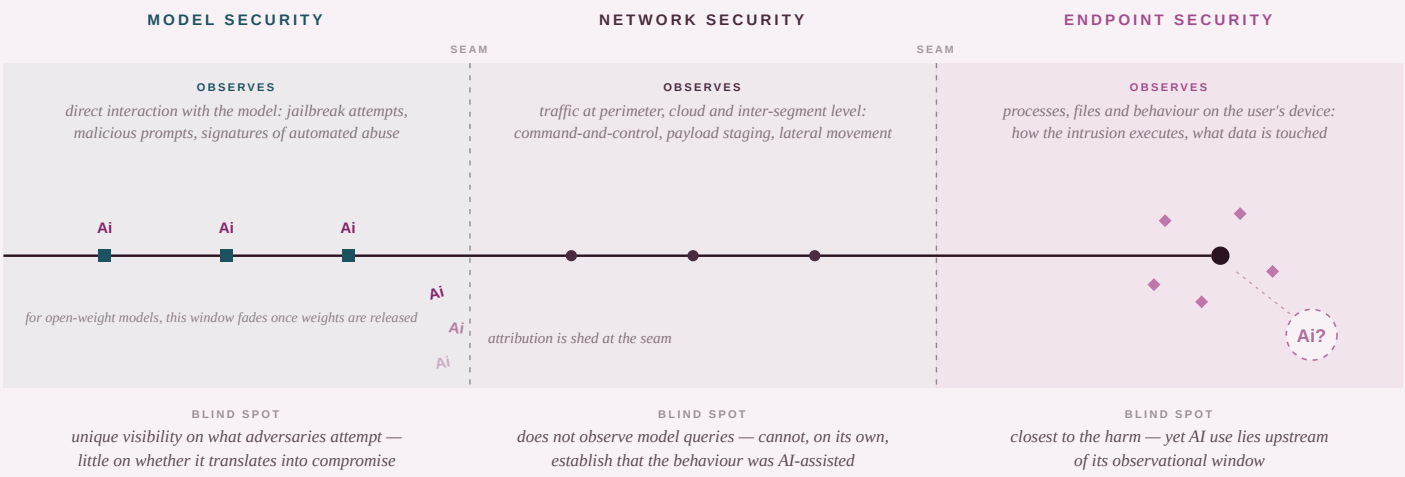
2.2 Sectoral fragmentation of telemetry

The evidence gap is compounded by a structural fragmentation of telemetry. No single class of actor observes more than a fraction of the AI-enabled attack chain, and the analytical categories used by each community to describe what it sees are not readily interoperable. Three principal vantage points can be distinguished.

- **Model security**. Frontier model developers observe direct interactions with their systems: jailbreak attempts, malicious prompts, agentic exploitation sequences, and the engineering signatures of automated abuse pipelines. They hold unique visibility on what adversaries attempt to make models do, but limited visibility on whether and how those attempts translate into downstream compromise. Open-weight developers, by contrast, lose visibility almost entirely once model weights are released; their telemetry fades from a real-time signal into coarse-grained retrospective inference.

- **Network security**. Network security providers analyse traffic at perimeter, cloud and inter-segment levels. Their visibility is essential to detect command-and-control infrastructure, the staging of payloads and the lateral movement that follows compromise. They do not, however, observe model queries themselves and cannot, on their own, establish that a given malicious behaviour was AI-assisted.

- **Endpoint security**. Endpoint security vendors detect suspicious processes, files, and behaviour on user devices. They are the closest observers of how an intrusion is executed and what data is touched, but they are typically unable to establish a causal link to AI use by the attacker, which remains upstream of their observational window.



Each layer provides a necessary but partial view, difficult to reconcile with the others. It is produced in different formats, on mismatched time scales, under divergent legal and classification regimes, and without a common vocabulary for describing events. The challenge is therefore not only to share additional data, but to align records that were never designed to interoperate.

Beneath the sharing problem lies a harder one of correlation. Even when the three layers are brought together, they may not answer the questions that matter most for policy: whether AI materially improved a phishing campaign, accelerated the development of an exploit, or raised the quality of a piece of malware. No shared event ontology yet exists to mark an incident as AI-involved across them. And in the moment of an intrusion, a defender rarely needs to establish that a payload was machine-generated; the immediate task is to detect and contain it. The value of bringing these vantage points together therefore lies less in attributing AI to any single case than in revealing how the threat landscape is shifting over time.

Furthermore, gauging how far AI actually changes an attacker's results, and who bears the harms, means cross-referencing the telemetry these layers produce with the further sources needed for aggregate analysis, from vulnerability and exploitation data to what the targeted sectors themselves observe. That combined picture can begin to show defenders where to harden first and governments which risks warrant a collective response.

A CIVIL SOCIETY PERSPECTIVE – **PROTECT.NGO**

"AI is not just accelerating cyber threats; it is reshaping who is exposed to harm. As attacks become more scalable, automated, and deniable, civil society organizations are increasingly targeted precisely because they lack visibility, resources, and protection. Without transparency on how AI is used in cyber operations, we risk building policies on incomplete evidence—leaving the most vulnerable outside the scope of protection and accountability."

Stéphane Duguin – CEO, Protect.ngo

2.3 A collaborative momentum yet to cross borders

The first half of 2026 has seen a notable acceleration of collaborative efforts at the AI-cyber intersection. Anthropic's Project Glasswing and OpenAI's Trusted Access Programme for Cybersecurity have both organized structured partnerships between frontier AI developers and cybersecurity practitioners, granting selected security providers early or privileged access to model capabilities for defensive purposes. These programmes represent a new category of public-private arrangement, distinct from both traditional threat-intelligence sharing and conventional vendor partnerships.

They build, moreover, on a broader foundation. Within the AI security and cybersecurity communities respectively, industry-led coalitions had already developed sustained mechanisms for pooling resources, sharing threat intelligence and aligning on methodology.

While this collaborative momentum is already producing tangible results, **the comprehensive picture on which credible international responses depend will require a wider geographic base**, extending structured participation to partners beyond a handful of jurisdictions. It will also require greater depth in what is shared, moving from high-level assessments toward more granular, comparable exchanges on the nature, frequency and impact of AI-enabled threats across sectors and borders.

Ultimately, a broader and deeper knowledge base must also reach those who can act on it. Closing the gap therefore turns as much on connection as on collection, and calls for effective channels that carry frontline insights into the forums where norms take shape.

AN INDUSTRY PERSPECTIVE – **WAVESTONE**

"The acceleration is no longer theoretical. In our incident response and threat intelligence engagements, we see AI compressing every phase of the attack chain, from reconnaissance to initial access to lateral movement. What used to require weeks of skilled tradecraft now unfolds in hours, sometimes with operators who could not have executed it a year ago. Yet the observation effort cannot rest on model providers alone. Most large enterprises we advise still have little to no monitoring of how AI systems are used within their own environments, which means they can neither detect misuse nor contribute meaningful signal to a collective evidence base. Both perspectives need to come together.

An international initiative that establishes a shared taxonomy and demonstrates the concrete value of exchanging observations will do more than inform policymakers. It will give every organization, regardless of maturity, a reason and a framework to start looking."

Gérôme Billois – Partner, Cybersecurity & Digital Trust, Wavestone

Part Three

An Architecture for a Shared Assessment

3.1 Introducing INTAiC

Since 2024, frontier developers and cybersecurity vendors have increasingly documented AI misuse in cyberspace. Each, however, stands on its own, and none was designed to combine into a picture that decision-makers can act on. To provide the empirical foundation on which an international consensus on AI-cybersecurity can rest, the next step must be a collective one. This requires an architecture that connects these siloed efforts while diversifying the contributors, sources, and analyses needed to achieve the full picture.

That architecture takes shape as INTAiC, the international initiative launched by the Paris Peace Forum to link real-world technical evidence to policy and diplomatic action. INTAiC brings together AI labs, cybersecurity firms, researchers, civil society, and public agencies to share trusted, sustained insight into how models are being misused and the impact of such misuses, making this data actionable for advancing an international response. Building on both the International AI Safety Report process and the Paris Call for Trust and Security in Cyberspace, this effort will be guided by five core principles:

- **Multistakeholder and impartial**. Industry, public agencies, international organisations and civil society take part on equal terms, with no single interest steering the outcome.

- **Globally inclusive**. The participation and engagement strategy extends well beyond the few jurisdictions where leadership on the AI-Cyber Nexus is concentrated today.

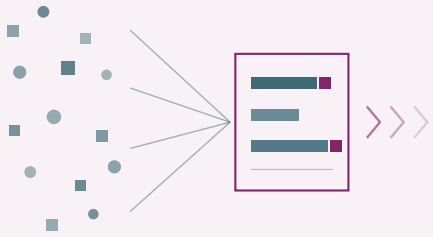
- **Shareable by design**. The coalition is not a real-time threat-intelligence exchange but a periodic synthesis of what partners already publish or choose to contribute, each keeping control of its proprietary data. Its value lies in the combined analysis, not in any single contributor's holdings.

- **Common vocabulary**. Brought together by the coalition, the AI security and cybersecurity communities gradually come to share a baseline vocabulary on AI misuse risks, so that their observations can be compared and combined rather than treated in isolation.

- **Calendar-anchored.** Outputs are timed to the international political agenda and disseminated across the major multilateral and sectoral milestones, so that findings reach decision-makers when they can be used.

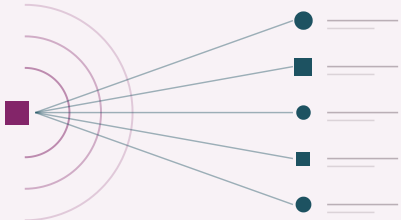
INTAiC takes its place in a wider field of work at the intersection of AI and cybersecurity, and is designed to build on it. It coordinates closely with the existing efforts in this space, including the Frontier Model Forum, the Coalition for Secure AI and the World Economic Forum's Cyber Frontiers initiative, whose contributions it seeks to extend and connect rather than to repeat. Because many of the organisations it brings together already take part in them, alignment comes naturally, and the coalition can add value where it is most needed.

3.2 Three lines of effort



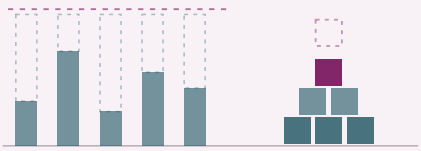
1 Map.

The AI-Cyber Threat Index, INTAiC's analytical core, draws together under a shared methodology the analysis of participating organizations on AI misuse in cyberspace. Rather than tracking isolated incidents, it aggregates analysis to assess the effectiveness of AI misuse, the capability it adds to attackers, and the consequences it produces, from the pressure it places on remediation to the burden carried by the least-protected targets. First released at the Paris Peace Forum in November 2026 and updated each year, the Index gives policymakers a single reference point and identifies the priorities that warrant collective attention in the short term.



2 Channel.

A sustained programme of briefings carries the INTAiC community's insights into the relevant international processes on AI and cybersecurity governance, to raise awareness and keep the issue both present and properly approached on their agendas, at a time when the track record can still look thin. Its role is to supply the shared evidence from which policy-makers can draw their own conclusions. Identified venues to date include the AI Summit Series, the United Nations' Global Mechanism on ICT security and its Global Dialogue on AI Governance, and the G7 and G20.



3 Build capacity.

The community also works with individual governments, strengthening how they confront AI misuse in cyberspace. Through tailored training and support, it backs the development of strategies and the policy capacity to act, engaging national cybersecurity agencies, AI safety and security institutes and ICT ministries. Particular attention will centre on the digital middle powers, where the ability to grasp the cyber threat landscape in the intelligent age remains uneven.

3.3 Resources to consider for a unified picture of the threat

01 Model-Side Abuse Telemetry

- Aggregated trends of malicious prompt injections.
- Automated API abuse patterns and safety-bypass (jailbreak) frequencies.

ANALYTICAL VALUE

Detects adversarial intent and early-stage weaponization at the source, before attacks hit downstream networks.

02 Downstream Behavioral Signals

- AI-mutated malware execution behavior (endpoint logs); velocity and success rates of automated phishing campaigns.
- AI-orchestrated attack infrastructure, including infrastructure-as-code provisioning and "living-off-the-land" activity.

ANALYTICAL VALUE

Measures the actual "capability uplift" — how much stronger, faster, and more resilient AI makes the average hacker.

03 Vulnerability Forecasting Data

- Daily EPSS (Exploit Prediction Scoring System) probability shifts.
- Machine-compressed timelines between patch release and exploit discovery.
- Active-exploitation signals from public catalogues.

ANALYTICAL VALUE

Shakes up traditional vulnerability management with a continuous, data-driven triage system for defenders.

04 Darkmarket & Underground Intelligence

- Commercialization of malicious LLMs (e.g., threat-specific models).
- Pricing and availability of automated hacking tools and stolen access on illicit forums.

ANALYTICAL VALUE

Maps the democratization of threats — how quickly sophisticated AI capabilities trickle down to low-tier criminals.

05 Aggregated Sectoral Impact Records

- Anonymized, cross-sector incident data from critical infrastructure hubs (ISACs).
- Impact metrics on under-defended targets (NGOs, hospitals, local governments).

ANALYTICAL VALUE

Focuses on the macro-level consequences of attacks on society rather than individual technical attribution.

Part Four

Engaging in the Initiative

Addressing the evidence gap on AI misuse in cyberspace is a systemic challenge that lies beyond the reach of any single organization. The Paris Peace Forum invites the multistakeholder community from both AI security and cybersecurity practice, spanning policy and technical circles, to actively engage in this collaborative framework.

THE VALUE OF CONTRIBUTING

i.

Enhancing Systemic Resilience

Participants leverage aggregated, cross-layer insights to keep pace with the use of AI by cyber adversaries. Accessing this collective intelligence allows organizations to proactively strengthen their own operational security posture and protect their broader ecosystems.

ii.

Advancing Expert Capabilities

Experts supporting the framework expand their capabilities through direct peer-to-peer learning. Working alongside world-class practitioners and experts accelerates institutional knowledge and advances cutting-edge competencies.

iii.

Driving Evidence-Based Policy

The initiative provides an impartial platform to facilitate trusted public-private dialogue. Members have a strategic opportunity to interface with policy makers and national authorities to inform international processes and foster global normative alignment.

JOIN THE COALITION

The Paris Peace Forum welcomes expressions of interest from organizations across AI security and cybersecurity practice, policy and technical circles alike.

References

[1] *AI and Attack Surface Complexity Fuel Majority of Breaches*, Palo Alto Networks – Unit 42, Feb. 2026.

[2] *2026 Global Threat Report*, CrowdStrike, Feb. 2026.

[3] *How Fast Is Autonomous AI Cyber Capability Advancing?*, UK AI Security Institute, May 2026.

[4] *Anticipating the Evolution of Critical AI Capabilities*, GPAI Policy Lab, Jan. 2026.

[5] *Cost of a Data Breach Report*, IBM, 2025.

[6] *The Emergence of Autonomous Cyber Attacks: Analysis and Implications*, Institute for AI Policy and Strategy, Nov. 2025.

[7] *Our Evaluation of Claude Mythos Preview's Cyber Capabilities*, UK AI Security Institute, Apr. 2026.

[8] *The Dual-Use Dilemma of AI: Malicious LLMs*, Palo Alto Networks – Unit 42, Nov. 2025.

[9] *2025 AI Threat Report: How Cybercriminals Are Weaponizing AI Technology*, KELA, Mar. 2025.

[10] *Threats to the Homeland: Cyber Operations Targeting US Government and Critical Infrastructure*, Check Point, Dec. 2025.

[11] *International AI Safety Report*, Feb. 2026.

[12] *Managing Advanced Cyber Risks in Frontier AI Frameworks*, Frontier Model Forum, Feb. 2026.

[13] *Challenges to the Monitoring of Deployed AI Systems*, U.S. NIST / Center for AI Standards and Innovation, Mar. 2026.

Acknowledgments

Disclaimer - The findings, interpretations, and conclusions presented herein are the outcome of a collaborative process facilitated and endorsed by the Paris Peace Forum. However, they do not necessarily reflect the exact views of each contributing organization or individual, or those of all member and partner organizations of the Paris Peace Forum.

Illena Armstrong
President, ***Cloud Security Alliance***

Gérôme Billois
Partner, Cybersecurity and Digital Trust, ***Wavestone***

Yann Bonnet
Deputy Chief Policy Officer, ***Paris Peace Forum***

Francesca Bosco
Independent Expert

Elsa Bouly
Diplomat, ***Ministry for Europe and Foreign Affairs of France***

Nicholas Butts
Director, Global AI and Cybersecurity Policy, ***Microsoft***

Lola Carbonell
Head of International Affairs, ***GPAI Policy Lab***

Pauline Charazac
Head of Policy Engagement, ***CeSIA***

Michael Daniel
President & CEO, ***Cyber Threat Alliance***

Stéphane Duguin
CEO, ***Protect.ngo***

John Hultquist
Chief Analyst, ***Google Threat Intelligence Group***

Paolo Palumbo
Vice-President, WithSecure Intelligence, ***WithSecure***

Paolo Pellegrino
Cybercrime Strategy & Capabilities Officer, ***INTERPOL***

Jean-Bernard Prouhet
Expert, ***Orange Cyberdefense***

Jim Ravis
CEO, ***Cloud Security Alliance***

Keir Reid
Research Affiliate, ***Oxford Martin AI Governance Initiative***

Wicus Ross
Senior Security Researcher, ***Orange Cyberdefense***

Pascal Steichen
CEO, ***Luxembourg House of Cybersecurity***

Adriana Stephan
AI Security Manager, ***Frontier Model Forum***

AN INITIATIVE BY THE PARIS PEACE FORUM

parispeaceforum.org