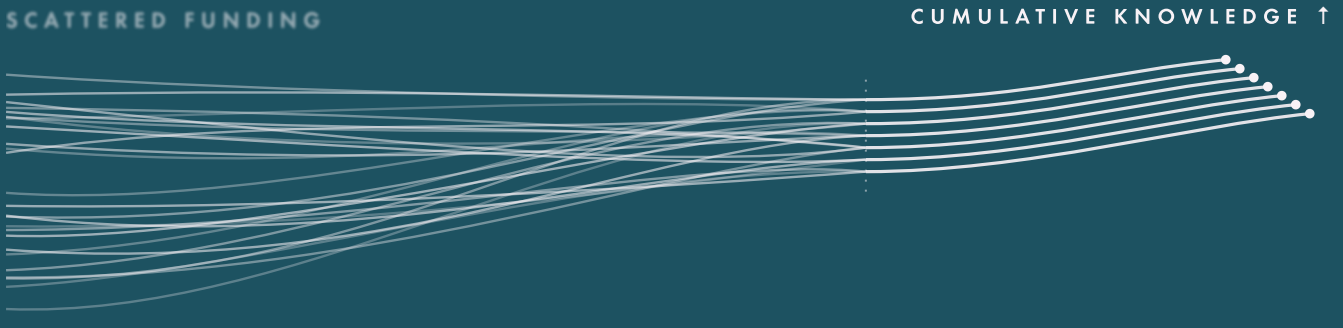
PARIS PEACE FORUM

INTAiC

Unlocking Public-Interest Research on Cyber-Capable AI

*Channeling investments to **anticipate** AI-enabled cyber threats*



INAUGURAL WHITE PAPER #2

JULY 2026

Foreword

This white paper opens the second 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, or could be, misused by adversaries in cyberspace. Where the initiative's parallel line of effort seeks to examine how such misuse – already occurring "*in the wild*" – can be reported, this one turns to the question that sits immediately upstream: who produces the public-interest research that anticipates what these systems make possible before misuse occurs, who funds that work, and why the current answer falls short of the need. Our critical infrastructure – from energy grids to hospitals, financial systems and communications – now rests on lines of code, and what it takes to attack that code is changing faster than anyone can track. On that front, we are still moving in fog.

It draws on a structured consultation process conducted jointly by the Paris Peace Forum and the Centre pour la Sécurité de l'IA (CeSIA) with researchers and funders across ten countries, opened by a thematic roundtable convened in Paris in February 2026, alongside the second annual conference of the International Association for Safe & Ethical AI. Their contributions, quoted anonymously throughout these pages, ground the paper's findings in the daily experience of those who conduct and support this research.

This effort does not duplicate what exists – it lifts the fog in which today's investments are made. It builds on the research priorities articulated by the scientific community, on the work of AI safety and security institutes, and on the programmes of the philanthropic and public funders discussed in these pages. It acts where no funder, agency, or lab can act alone: aligning fragmented investments with the questions policymakers most need answered.

As the workstream's co-leads, the Paris Peace Forum and CeSIA are grateful to every respondent and participant whose insight informed this document. INTAiC remains open to all stakeholders who share the conviction that trusted evidence on AI-enabled cyber risk is a public good worth funding together.

Charbel-Raphaël Segerie, Executive Director, CeSIA

Pablo Rice, Head of Programme, Paris Peace Forum

Executive Summary

This white paper examines what keeps public-interest researchers, and the funders behind them, from anticipating AI-enabled cyber risks early enough for governments to act. We ask how investment in independent research at the intersection of AI security and cybersecurity is currently distributed, where the structural obstacles lie, and what those obstacles mean for aligning investments and organising research programmes more effectively.

The paper approaches the evidence gap in AI-cybersecurity from the angle of research infrastructure. In under a month, Claude Mythos identified more than a thousand serious vulnerabilities in the software we all use every day — some of which had gone unnoticed for more than twenty years despite rigorous human review [1]. A year ago, such discoveries were the preserve of a handful of states and highly specialised groups. That scarcity is ending — and the research that would tell governments what comes next is nobody's budget line. Research investment at this intersection remains fragmented across governments, philanthropies, and borders. No shared map exists of what is funded, what is not, and where the most consequential gaps lie. Established funding instruments were not designed for a field that needs both long-term capacity and answers within months.

Independent evaluation of AI cyber capabilities, in particular, emerges as a precondition for any credible governance framework. For most governments, the exposure is double: their infrastructure faces threats born of technologies developed beyond their borders, and they depend on those same technologies — and on their developers' goodwill — to understand the risk, let alone to defend against it. Findings cannot be compared, challenged, or carried forward when every team evaluates on its own benchmarks, its own model access, and its own compute. **The constraint this paper identifies is not so much a shortage of money as an architecture of allocation**: for every four hundred dollars invested each year in advancing AI capabilities, roughly one goes to philanthropically funded AI safety and security research [2] — and almost none of that reaches the cyber intersection.

The paper closes with what this workstream proposes to do about it. The consultation that grounds these pages has opened a standing dialogue between researchers and the funders who support them. **The next step is turning this dialogue into a joint research roadmap to guide public and philanthropic funding toward shared goals**. The first results of that work will be brought to the ninth Paris Peace Forum in November 2026. The priorities it serves are being defined now, and they will bear the mark of those who help define them. AI can become an instrument of cyberdefense at least as powerful as of attack. It will not happen by default — and neither will the evidence base that makes it possible.

Part One

A field divided across research communities

Artificial intelligence has the potential to reshape the cyber risk environment at a pace that tests the tracking capacity of even the best-resourced governments. Independent evaluations conducted by national AI safety and security institutes find that the autonomous cyber task horizon of frontier models, that is, the length of cyber tasks they can complete without human intervention, was doubling every 4.7 months as of early 2026, an acceleration from the eight months estimated in late 2025 [3]. Anticipating what these systems make possible is therefore a standing task. It requires researchers able to work across fields and institutions where that work can live. This Part examines how far those conditions are met today.

1.1 The missing foundation for knowledge production across communities

"AI-enabled cyber risks" in practice span three distinct disciplines that different communities hardly treat as one. AI safety is concerned with how systems behave: helping ensure that they act ethically, reliably, fairly, transparently, and in alignment with human values. Misuse – a person directing a model's capabilities toward unauthorised ends – engages this discipline first, and it supplies the defences that matter there, from interpretability to jailbreak resistance, though none of them ever closes the gap completely. AI security is concerned with protecting the system itself: guarding AI systems against unauthorised use, availability attacks, and integrity compromise across their lifecycle, and keeping increasingly autonomous systems accountable – an extension of the cybersecurity principles of confidentiality, integrity, and availability, applied to assets such as model weights and training infrastructure. Classical cybersecurity, finally, treats AI as one link in an otherwise familiar attack chain: a risk of degree rather than kind. Misuse cuts across all three: it exploits model behaviour, targets the systems themselves, and feeds otherwise ordinary intrusions. This is why no single community sees the whole of it, and why safety and security are best read as complementary dimensions of a single risk picture [4].

These disciplines combine less than their complementarity would suggest, not least because the communities that practise them are organised apart. Each community therefore sets its priorities in isolation, around the questions its own discipline already recognises. Work that crosses both seldom finds an institutional home or a dedicated budget line. Left unaddressed, this channelling risks sealing each community into an increasingly closed circle: the evidence produced never accumulates, scattered across environments that cannot reproduce one another's results. This divide is the first of the structural obstacles this paper identifies, and it sets the pattern for those that follow.

THE 2026 SINGAPORE CONSENSUS

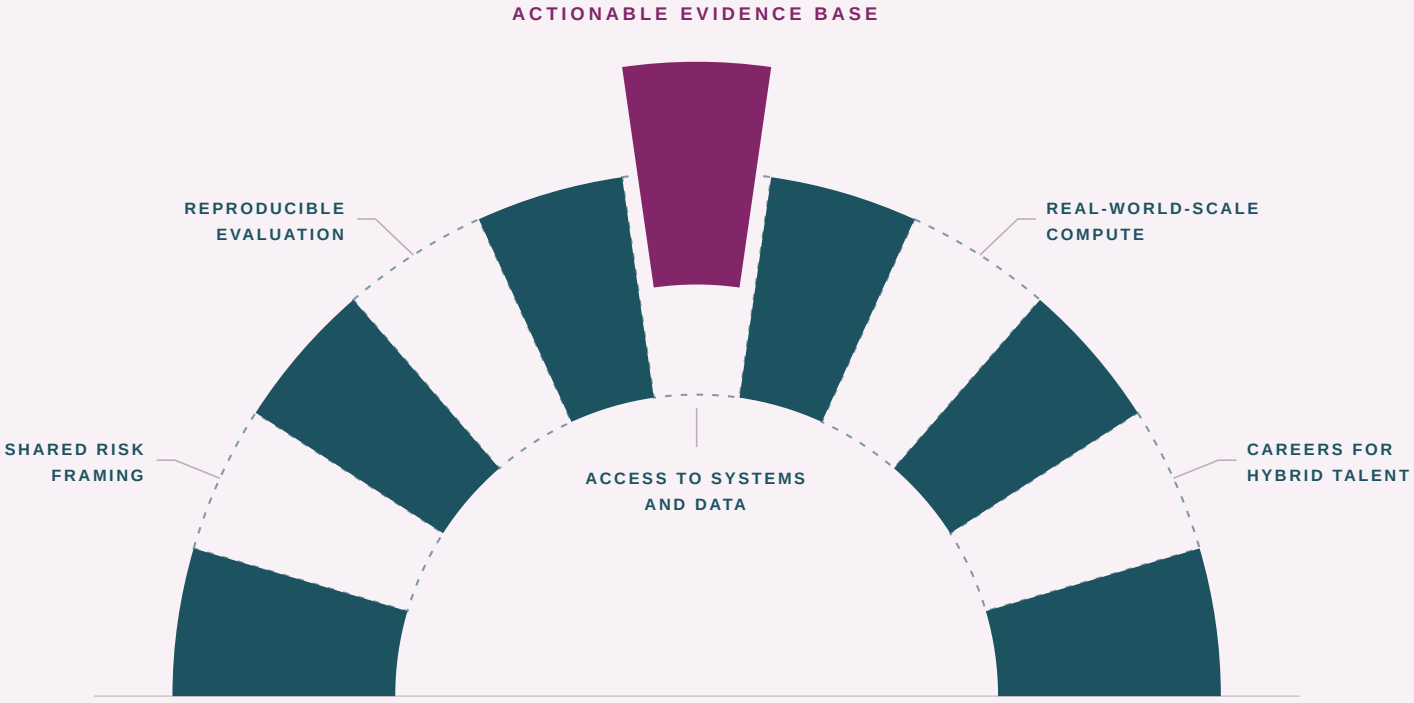
AN AI SAFETY COMMUNITY'S RENEWED FOCUS ON CYBER MISUSE

In May 2026, more than one hundred researchers from thirteen countries convened in Singapore for the second International Scientific Exchange on AI Safety. The resulting **2026 Singapore Consensus on Global AI Safety Research Priorities [5]** updates the community's first shared research agenda for a landscape it describes as reshaped by autonomous agents and by misuse incidents that "*have grown precipitously, particularly for cyber attacks*."

Prompted by what the report calls an abrupt increase in cyber-relevant model capabilities, the 2026 edition moves cyber misuse to the front of the risks it puts to policymakers, ahead of the biosecurity and child-safety concerns it also treats. It asks for cyber-capability evaluations tough enough to keep pace with frontier systems, and for tools that tilt the advantage toward defenders. It looks to the monitoring of coordinated misuse across deployed agents, and to shared reporting of serious AI-enabled cyber incidents. These are the same research tasks whose practical conditions the rest of this section examines.

1.2 Practical Obstacles to Meaningful Synergies

Even where the two communities set out to work together, practical obstacles stand between them and a usable evidence base. The consultation this paper draws on asked researchers from both the AI safety/security and the cybersecurity communities what holds back substantive work at their intersection. What they describe rests on structural conditions that remain only partly met today – not on shared goals alone.



1. No shared approach to modeling risks and threats.

The two communities do not define their object the same way. Cybersecurity treats AI as a new degree of a familiar problem, an adversary exploiting weaknesses in assets it can already describe, and reasons from incidents on record. The AI safety and security community treat it as a new kind of problem centred on the model itself, and reason ahead of the evidence about capabilities not yet seen. Frameworks that bring the two together are appearing, but they organise the defence of AI systems and say little about how much AI would change what an attacker can do. On that question no shared baseline yet exists, so the same development is weighed differently or missed.

2. Evaluations that cannot yet be reproduced.

Shared tests of what models can do in cyber tasks have appeared and are being taken up across institutions. They still rest largely on stylised exercises that models soon master, and no agreed standard defines what counts as adequate evidence, so results are hard to compare or to verify independently.

3. Access to systems and data on discretionary terms.

Access to the most capable models, and to the behavioural data that shows how they act under adversarial testing, has improved mainly for a small circle of national bodies and approved evaluators. Some jurisdictions are starting to make that access compulsory, though on terms that channel it to regulators and the evaluators they designate. For the wider independent research community it still rests on arrangements developers grant and can withdraw. Public-interest researchers remain dependent on the very actors whose systems they would assess, for both the systems and the evidence about them.

4. Compute to test under real-world conditions.

Testing models under conditions close to real operations now demands heavy compute. Measured capability keeps rising with the compute an evaluator can spend, so those with less will understate what a system can do. Public compute for research is growing, but it stays a fraction of industry's, and little of it is set up for this testing. Independent evaluation is then left reporting the limits of its budget as the limits of the system.

5. A hybrid expertise that careers do not sustain.

The work needs people who combine machine-learning and operational security skills, a rare mix that current incentives pull toward well-resourced laboratories offering higher pay and better computing. Funding and publication still reward staying within one discipline, so the expertise on which every other obstacle depends is the hardest to build and to keep.

CONSULTATION RESPONDENT – **INDEPENDENT RESEARCHER, SWITZERLAND**

"We already broadly know what we need to study; what we lack is the shared, governed substrate that makes those studies reproducible, comparable, and safe to conduct at scale."

Part Two

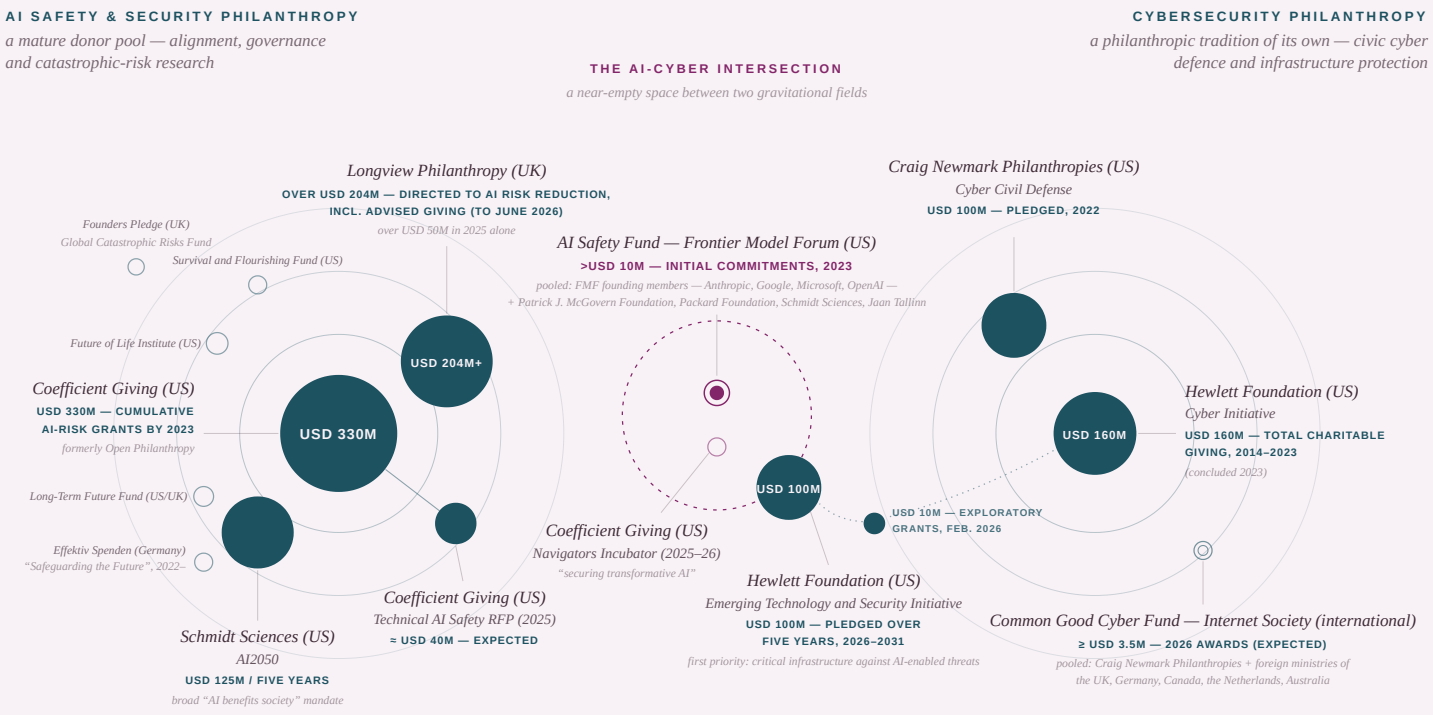
An investment architecture in need of realignment

Closing the evidence gap through public-interest research is often framed as a question of how much is spent. What the record assembled here points to is the architecture through which the money moves, built inside two separate fields and calibrated to their horizons.

2.1 Mapping the current funding base

Philanthropic funding

The philanthropic ecosystem began supporting research on the misuse risks associated with advanced AI well before institutional funders, through sustained support for the AI safety and AI security communities. Until the first national AI safety institute was established by the United Kingdom in late 2023, AI safety research was a niche field, funded by a small circle of specialised donors and a few forward-looking programme officers. The picture has shifted, while remaining predominantly private and philanthropic rather than public. Open Philanthropy, since renamed Coefficient Giving, had cumulatively administered some USD 330 million in grants to the field by 2023 [6], and its largest open call to date substantially exceeds what any single U.S. federal open call has directed toward AI safety specifically. The wider pool of US foundation giving to AI runs to several hundred million dollars over the past five years. Within these totals, the share devoted to AI-cyber risks was until recently barely identifiable. In July 2026, however, the Hewlett Foundation announced a five-year, USD 100 million Emerging Technology and Security Initiative [7] which, though broader in scope, makes the protection of critical infrastructure against AI-enabled threats a core priority.



Continental European and Japanese philanthropy engages AI mainly through policy and civil-society programmes (Stiftung Mercator, Robert Bosch Stiftung, the pooled European AI & Society Fund, Sasakawa Peace Foundation) — none fund technical AI safety or security research at scale.

Selected disclosed commitments — illustrative, not exhaustive. Amounts differ in nature (cumulative grants, pledges, multi-year programmes, single calls) and period; each label states its basis. Bubble areas are proportional to disclosed amounts and compare scale, not like-for-like flows.

Cybersecurity research has been funded the other way round: public budgets and industry have carried the technical field, while philanthropy played a narrower but catalytic part, building its civic and policy dimension. Government agencies and programmes remain the dominant funders of cybersecurity research across the G7, through programmes of a scale philanthropy has never approached. The establishment of the Hewlett Foundation's Cyber Initiative in 2014 was itself a response to the near-absence of philanthropic funding for cybersecurity. Over the following decade, its USD 160 million made it, in the words of one observer of the sector, among the only funders of any consequence in the field. That period ended in 2023 when the initiative concluded as planned, leaving a gap practitioners were quick to name [8]. What remains is more modest in scale. Craig Newmark Philanthropies' Cyber Civil Defense commitments are directed principally at protecting critical infrastructure and communities in the United States; the pooled Common Good Cyber Fund, launched in 2025, was created to sustain the nonprofit organisations whose work strengthens the public-interest cybersecurity ecosystem. Research at the intersection currently benefits from neither: public cybersecurity programmes rarely extend to it, and philanthropic support for AI safety seldom reaches it.

The knowledge base for AI safety and security rests on a small number of funders, most of them based in the United States and the United Kingdom. Most G7 national ecosystems do not fund this research themselves; the evidence their policies draw on is produced and financed elsewhere. This concentration comes with little redundancy. When a funder withdraws, no other source stands ready to take over its grantees. The collapse of the cryptocurrency exchange FTX in 2022 illustrated that exposure: the research fund it financed, which had committed some USD 30 million to AI safety projects in its nine months of existence, disappeared with it, and grantees lost their support overnight, in the United States first.

Public funding

Government engagement with the AI-cyber intersection is recent, and accelerating. The national AI safety institutes created from late 2023 onwards gave governments their first standing capacity to evaluate the cyber capabilities of frontier models, and dedicated funding instruments followed.

- **Germany**'s BMBF opened a first call on AI and IT security in 2024, framed as applied security technology. In June 2026 its National Security Council approved a dedicated AI security institute built on the Federal Office for Information Security and the Federal Network Agency, spanning model evaluation and operational cybersecurity. It develops evaluation capacity without, so far, funding external research. [9]

- **France**'s INESIA, launched in January 2025, brought the national cybersecurity authority together with AI and metrology bodies under an explicit remit, though its external grant architecture was not specified at the time of writing. Its February 2026 roadmap made the intersection an explicit priority, naming offensive cybersecurity among the systemic risks it will study and committing the institute to evaluation and certification methods for the security of AI systems. [10]

- The **United Kingdom** paired its AI Security Institute with the National Cyber Security Centre and has built one of the most developed public grant portfolios in AI safety, several schemes open to UK and international researchers. The largest of these, the Alignment Project, assembled the first pooled international fund of its kind, though its opening round has closed, with further academic programmes expected.

- In the **United States**, the evaluation institute renamed the Center for AI Standards and Innovation in 2025 works through an interagency taskforce and has assessed dozens of frontier models, including their cyber capabilities, some in classified settings. A June 2026 executive order added a classified National Security Agency benchmarking process for frontier models' cyber capabilities, with findings shared only as appropriate. Under it the Gold Eagle clearinghouse launched that July to coordinate vulnerability remediation across critical infrastructure [11]. This national-security strategy, narrowed under the current administration, evaluates and defends; it does not fund outside research. Open, university-led work at the intersection is funded separately, chiefly through the National Science Foundation, though that base has come under budget pressure.

- **Canada**'s main vehicle is a partnership between its research council and its signals-intelligence agency, which co-funds unclassified, publishable research explicitly framed around robust, secure and safe AI, in multi-year projects worth several million dollars each. This sits alongside a national AI safety institute launched in 2024, whose broader research programme of some twenty-seven million Canadian dollars addresses AI safety more widely, and a June 2026 national strategy renewed the government's commitment. [12]

- **Italy** passed the European Union's first national AI law in 2025, earmarking up to one billion euros for AI and cybersecurity and naming its national cybersecurity agency as the authority for AI in the cyber domain. That money flows to companies through venture capital rather than to open research, and the agency's role at the intersection is regulatory. [13]

- **Japan** built evaluation capacity without a funding arm. Its December 2025 AI strategy, revised in June 2026 as frontier threats grew, widened its AI safety institute's remit to evaluate advanced models and to guard critical infrastructure against AI-enabled attacks. The institute runs model evaluations and red-teaming, and in January 2026 opened those exercises to domestic security firms, but it commissions no external research. [14]

The public turn toward the AI-cyber intersection has therefore a geography, with a significant predominance from North America and the United Kingdom. The European Commission's Action Plan on Cybersecurity and Artificial Intelligence, published in July 2026 [15], takes this asymmetry as its starting point: it observes that frontier capabilities are mainly developed outside the EU, that most third-party evaluation of advanced AI models takes place outside the EU, and that access to advanced capabilities is granted through provider-specific processes over which European organisations have little visibility. Its response runs along three announced lines. On evaluation, it announces an EU evaluation capacity for AI models that must include cybersecurity, and criteria intended to foster a wider ecosystem of third-party evaluators. On access, it plans a European Blueprint for structured access to advanced AI capabilities, including contingency measures should access to a model be restricted or withdrawn. On funding, it earmarks a secure testing platform for cybersecurity use cases, dedicated research support under the Horizon Europe and Digital Europe programmes, a first SecureAI topic on the robustness of models against adversarial attacks, and a Grand Challenge on AI-assisted vulnerability remediation. Whether these instruments will support independent research, rather than regulatory evaluation and operational deployment, was not settled at the time of writing.

CONSULTATION RESPONDENT – **INTERNATIONAL ORGANISATION, FRANCE**

"Benchmarks and datasets are dominated by Global North and English-language contexts, so they systematically under-measure risks elsewhere."

2.2 The agenda-setting power of funding channels

How research is funded shapes what gets researched. Funding instruments carry assumptions about the horizon of a project, the discipline that will assess it and the form its results should take, and those assumptions select among the questions a field could pursue. The recent history of AI safety / security research illustrates the effect. The field grew from a narrow base after 2016 largely on the strength of one foundation's sustained commitment, which allowed a single set of strategic preferences to be expressed across fellowships, academic grants and new organisations at once. The same logic operates at the scale of the field as a whole. Of the AI models identified as notable in 2025, one originated in academia and eighty-seven in industry, and questions requiring frontier-scale resources are in practice asked where those resources sit.

Work that spans disciplines risks being disadvantaged by these arrangements. Proposals succeed less often as they reach further across fields, and that disadvantage weakens where assessment is not organised by discipline, which suggests a penalty attaching to how work is framed and routed rather than to the substance of crossing fields [16]. Earmarked calls correct this only partly. Thematic programmes shift research direction gradually, and by margins comparable to untargeted funding [17]. Money entering a field tends to reinforce the collaboration networks already present within it more than it widens the range of questions those networks pursue.

These effects compound at the AI-cyber intersection, as the consultation conducted for this paper records directly. Researchers describe funding opportunities framed around either technical AI development or cybersecurity operations, and proposals assessed against criteria that do not match the problem being studied. They report a misalignment of expertise in the evaluation of interdisciplinary projects, review processes that reward predictable outcomes and established methodologies, and work judged too applied for research funding while remaining too immature for implementation funding. Respondents describe adapting their programmes to the priorities available rather than to the gaps they consider most consequential. Funders describe the other side of the same arrangement, coordinating with one another on an ad hoc basis, each working from its own theory of change.

Most of the instruments mapped in the previous section were built within one field, and a call opened inside one of them reaches the community that field already contains. This points to a question of routing more than of volume.

CONSULTATION RESPONDENT – **PUBLIC AGENCY REPRESENTATIVE, TÜRKIYE**

"A misalignment of expertise in the evaluation of interdisciplinary projects often leads to the dismissal of high-potential research."

2.3 Allocation agility as a collective problem

The cadence of a competitive call is a structural feature of the system it belongs to. Across research funding systems the dominant frequency of calls is annual, and most schemes assess proposals through several successive stages, from initial screening to external review and panel deliberation [18]. That sequence is what gives an award its legitimacy, since judgement by peers is what makes public allocation defensible to the community it serves. It also takes time. Where success rates are low, the effort a research community spends competing can approach the value of the funding distributed [19].

CONSULTATION RESPONDENT – **PHILANTHROPY REPRESENTATIVE, UNITED STATES**

"We don't move as quickly as we would like. Best case scenario is likely five months from initial call for proposals to disbursement of funds. […] The Board review and approval is a bottleneck. And our payment process, especially for non-US-based organizations, is another bottleneck."

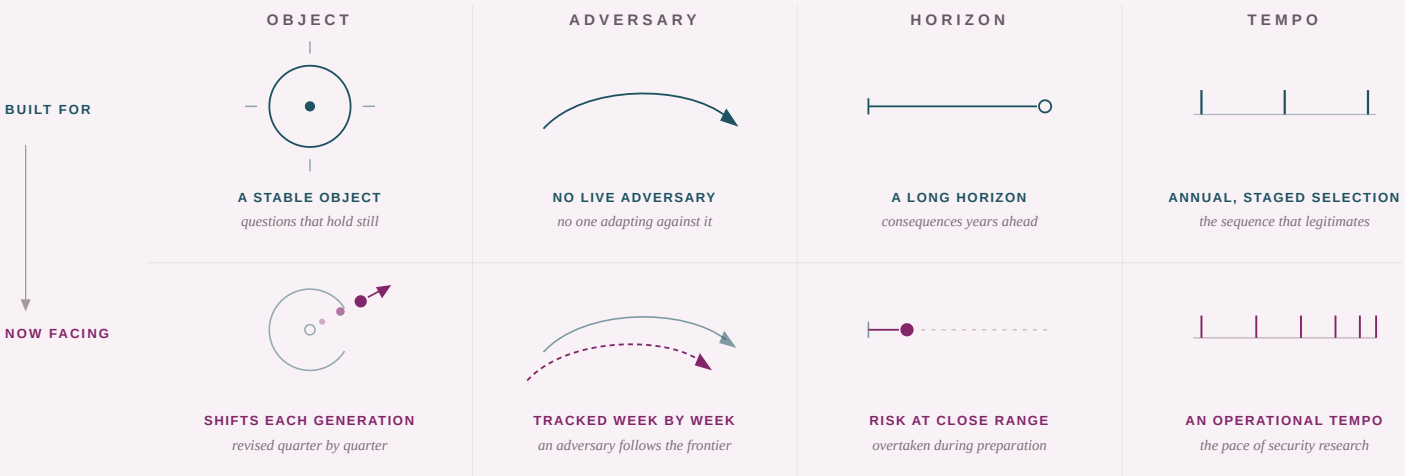
In a domain where transformative developments occur every quarter, that cadence carries a cost that is harder to absorb. Respondents describe threat scenarios that become obsolete during project preparation, and evaluation baselines overtaken before results are finalised. The pace has proved difficult even to measure. The estimated doubling time for the length of cyber tasks frontier models complete unaided was revised downwards within three months of the figure it replaced an interval shorter than most funding decisions take.

The mismatch runs deeper than tempo. The funding base on which AI safety research was built formed around questions with no live adversary adapting against them week by week. The instruments it still relies on were not built for the field, and they are better suited to long-horizon investigation of consequences that lie ahead. Anticipating AI-enabled cyber risk moves on the calendar of an adversary, not a grant cycle. Cyber capability evaluation and misuse threat modelling take as their object a capability frontier that shifts with each model generation, and an adversary presumed to be tracking it. A threat model scoped against one generation of systems can be superseded before the work assessing it is complete. The work therefore carries the tempo of operational security research rather than that of the agenda the field has been funded to pursue. Several consultation respondents observe that the long-term AI risk funding space rewards ambitious novel research while overlooking established norms in adjacent fields that could be adapted to it.

Shortening the cycle would not on its own resolve the difficulty, since the judgement it compresses is itself poorly reproducible. Beyond a summary verdict, the chance that two reviewers reach the same evaluation of the same proposal is close to random [20], and the tendency toward caution grows as success rates fall. Funders in other research fields have tested remedies, none of which has become standard practice. Some programmes reserve a fixed share of their funding for proposals judged genuinely high-risk, insulating them from competition against safer work. Wide divergence among reviewers' scores can be treated as a signal of promise, on the premise that novel proposals divide assessors where incremental ones do not. Allocation by lot among applications that reviewers cannot reliably separate has been used by one European foundation after a quality threshold [21]. Where proposals span disciplines, panels drawn from different fields can be required to coordinate a single joint evaluation.

These remain untried in research funding at the AI-cyber intersection, so far as published information shows. A field working to make the evaluation of AI systems reproducible and comparable is itself funded through a process that meets neither standard. Testing an alternative would require funders willing to depart from established practice together, since a mechanism adopted by one of them alone changes little about where the field's proposals are judged.

Agility lags by design



Part Three

A roadmap toward a common enterprise

3.1 Introducing INTAiC

In recent years, funders in AI safety and security have committed growing sums to research in their own field. So have funders in cybersecurity. Each has supported public-interest research on its own terms and on its own horizon, and the questions that fall between the two have hardly been anyone's remit. To meet them, **the next step must be a collective one, set to complement each funder's own theory of change while directly reflecting the thematic priorities and capacity needs expressed jointly by the research communities involved**.

This collective endeavor is what INTAiC's second pillar is designed to catalyse. Launched by the Paris Peace Forum to link real-world technical evidence to policy and diplomatic action. INTAiC brings together researchers and funders working in the public interest to converge on the priority work most worth pursuing at the AI-cyber intersection and to reflect on innovative arrangements to deploy such cross-disciplinary projects. 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 the following core principles:

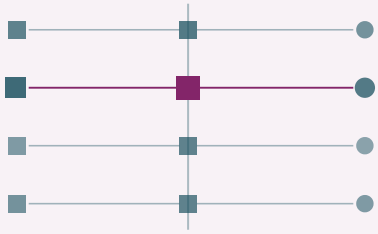
- **Multistakeholder and impartial**. Public agencies, philanthropies, international organisations, the industry, the civil society and academia take part on equal terms, with no single interest steering the outcome.

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

- **Common vocabulary**. Brought together by the coalition, the AI safety / 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.

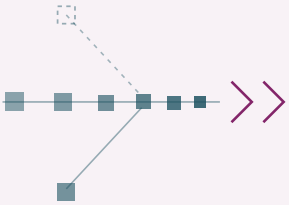
- **Calendar-anchored**. Outputs are timed to the international political agenda and disseminated across the major multilateral and sectoral milestones, so that findings reach decision-makers when they can be used.

3.2 Three lines of effort



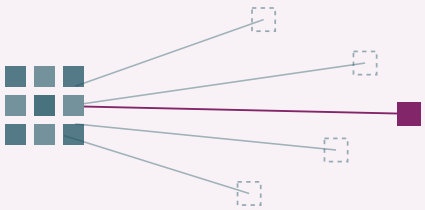
1 Align priorities.

The AI-Cyber Research Priorities Roadmap, the pillar's first output, draws the consultation reported here into a shared view of where research at the intersection should concentrate. Rather than the work of researchers alone, it is agreed by funders and researchers together, so that it connects the setting of priorities to the allocation of resources. First presented at the Paris Peace Forum and revised as the field moves, the Roadmap gives both communities a common reference on issues that warrant collective attention in the near term.



2 Adapt funding.

Alongside the Roadmap, **funders and researchers work on the conditions for funding cross-disciplinary projects that are both timely and relevant, from the pace of allocation to the breadth and consistency of the judgement that assesses them.** The same effort looks for where a commitment in one channel can unlock or extend one in the other. It seeks a shared practice each funder can adopt within its own mandate.



3 Distribute globally.

INTAiC community also carries this work into the international processes where AI and cybersecurity governance are shaped. Through them, it can contribute to a renewal of ICT capacity building attuned to how AI is reshaping cyber risk. Its aim is that the capacity to anticipate and evaluate AI-enabled cyber risk should reach well beyond the few jurisdictions that concentrate it today. Particular attention will centre on the regions where the consequences of underestimating these risks would be gravest.

Part Four

Engaging in the Initiative

Building the cumulative knowledge needed at the AI-cyber intersection lies beyond the reach of any single funder or discipline. The Paris Peace Forum invites funders and researchers from both AI safety / security and cybersecurity circles, together with the public authorities that draw on their work, to take part in INTAiC's second pillar.

THE VALUE OF CONTRIBUTING

i. **Extending the reach of each investment**

Funders gain a view of the whole field and a place to set their commitments beside those of their peers, each keeping its own mandate. From there they build on one another instead of overlapping, and together reach further than any funder could alone.

ii. **Backing work that falls between fields.**

Researchers gain a hand in shaping the priorities that guide funding, and practices better suited to work that moves quickly and crosses disciplines. Proposals that today fall between established programmes find a readier route to support.

iii. **Mainstreaming anticipatory readiness**

Public authorities gain a hand in making the anticipation of AI-enabled cyber risk a standard part of capacity building, rather than a specialism confined to a few. That gives even the least-resourced among them independent knowledge and evaluation capacity to draw on, and a voice in how the agenda adapts to AI.

JOIN THE COALITION

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

References

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

[2] ***2026 AI Index Report***, Stanford HAI, 2026. The 1:400 ratio compares global investment in AI capability development (of the order of hundreds of billions of dollars per year, per the AI Index) with annual philanthropic funding for AI safety and security research (of the order of two hundred million dollars; see [6]).

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

[4] ***Cisco Integrated AI Security and Safety Framework Report***, A. Chang et al., arXiv:2512.12921, December 2025.

[5] ***The 2026 Singapore Consensus on Global AI Safety Research Priorities***, Second International Scientific Exchange on AI Safety, May 2026.

[6] S. Campos and D. S. Schiff, "***The AI Extreme Risk Mitigation Philanthropic Sector***", in The Routledge Handbook of Artificial Intelligence and Philanthropy, Routledge, 2025.

[7] ***Hewlett Foundation Announces New $100 Million Emerging Technology and Security Initiative***, William and Flora Hewlett Foundation, July 2026.

[8] ***Marking the End of the Cyber Initiative***, William and Flora Hewlett Foundation, 2023; ***Findings from the Hewlett Cyber Initiative Summative Evaluation***, October 2023.

[9] ***German National Security Council Sets Up AI Security Institute***, Telecompaper, June 2026.

[10] ***Une feuille de route pour l'Institut national pour l'évaluation et la sécurité de l'IA (INESIA)***, ministère de l'Économie, 2026.

[11] ***Executive Order, Promoting Advanced Artificial Intelligence Innovation and Security***, Federal Register, June 2026; ***White House Launches Gold Eagle Initiative***, July 2026.

[12] ***New NSERC-CSE Partnership to Support Robust, Secure, and Safe Artificial Intelligence Research***, Natural Sciences and Engineering Research Council of Canada, 2023.

[13] ***Legge 23 settembre 2025, n. 132 (disposizioni in materia di intelligenza artificiale)***, Gazzetta Ufficiale, 2025.

[14] ***Factsheet of AI Safety in Japan 2025***, Japan AI Safety Institute (J-AISI), 2025.

[15] ***EU Action Plan on Cybersecurity and Artificial Intelligence***, European Commission, July 2026.

[16] L. Bromham, R. Dinnage and X. Hua, ***Interdisciplinary Research Has Consistently Lower Funding Success***, Nature 534, 2016.

[17] ***Does Targeted Research Funding Reshape Research Landscapes?***, Research on Research Institute, 2026.

[18] ***Effective Operation of Competitive Research Funding Systems***, OECD Science, Technology and Industry Policy Papers No. 57, 2018.

[19] D. L. Herbert et al., ***The Impact of a Streamlined Funding Application Process on Application Time***, BMJ Open, 2015.

[20] E. L. Pier et al., ***Low Agreement Among Reviewers Evaluating the Same NIH Grant Applications***, PNAS 115, 2018.

[21] ***Partially Randomized Procedure: Lottery and Peer Review (Experiment!)***, Volkswagen Foundation.

Lead Authors

Niharika Chaubey
AI Policy Fellow
CeSIA

Pablo Rice
Head of Programme
Paris Peace Forum

Aknowledgments

Charbel-Raphael Segerie
Executive Director
CeSIA

Francesca Bosco
Researcher
Independent

Pauline Charazac
Head of Policy Engagement
CeSIA

Stéphane Duguin
CEO
Protect.ngo

Yann Bonnet
Deputy Chief Policy Officer
Paris Peace Forum

Keir Reid
Research Affiliate
Oxford Martin AI
Governance Initiative

AN INITIATIVE BY THE PARIS PEACE FORUM

parispeaceforum.org