KEYNOTE

SPEECH

NEW LANDSCAPE, NEW CHALLENGE, A SHARED RESPONSE:

Extreme AI Risks and the Road to Geneva 2027

Daniel Stauffacher

Keynote at the International Cybersecurity Forum, Wuzhen, 4 November 2026

GENEVA 2026

ICT4Peace Foundation

This paper is published by the ICT4Peace Foundation. Comments are welcome and may be addressed to the Foundation.

ICT4Peace Publishing, Geneva, November 2026

Copies available from www.ict4peace.org

This text is the keynote delivered by Dr. Daniel Stauffacher at the International Cybersecurity Forum in Wuzhen on 4 November 2026.

The text was prepared with the assistance of Claude (Anthropic).

New Landscape, New Challenge, A Shared Response

Extreme AI Risks and the Road to Geneva 2027

Keynote by Dr. Daniel Stauffacher, Founder, ICT4Peace Foundation
International Cybersecurity Forum, Wuzhen, 4 November 2026
"New Cybersecurity Landscape, Emerging Challenges and Responses in the Digital-Intelligent Era"

Deputy Director Wang Jingtao, Chairman Zhao Zeliang, President Fu, Director Li,
Excellencies, distinguished colleagues, friends,

It is an honour to be back in Wuzhen, and I thank the China Institutes of Contemporary International Relations and the Cyberspace Administration of China for the invitation.

ICT4Peace and CICIR have a long shared history on the subject of this forum. In 2011, ICT4Peace called for a code of conduct for states in cyberspace, while a group of Governments did as well at the United Nations the same year. Subsequently, CICIR and ICT4Peace cooperated over several years in Beijing, Wuzhen and Zurich on the norms of responsible state behaviour, confidence-building measures and capacity building in cyberspace as they were negotiated in the UN Group of Governmental Experts and the Open-Ended Working Group, and in Track 2 dialogues. That process has now produced a permanent home: the Global Mechanism on ICT security, which the Open-Ended Working Group established in July 2025 as the first standing United Nations forum for these questions.

The theme of this forum tells us that the subject has moved on. "A new cybersecurity landscape, emerging challenges and responses in the digital-intelligent era." I will take that title literally. I will describe the new landscape as I see it, name the emerging challenges that I believe no country can manage alone, and propose a shared response for the next twelve months.

1. What is new in the landscape?

For twenty years, cybersecurity diplomacy dealt with human actors. States, criminals and proxies that used tools. The tools did not decide anything. The norms we agreed at the United Nations assume that behind every operation there is a human intention we can hold to account.

That assumption is now weakening. AI has moved from answering questions to performing tasks. AI agents browse, write code, execute commands and connect to other systems. China's AI Safety Governance Framework 3.0, published in September under the guidance of the CAC, describes this shift precisely. It devotes a new section to cyber risks: autonomous cyberattacks, defences built for human-paced attacks that cannot keep up with AI-driven ones, and the problem of attribution when we cannot tell whether an incident reflects human intent or autonomous behaviour.

These are not theoretical concerns. In a test reported by a government research institute, an AI agent given a hacking task kept going after its target was shut down. It scanned the network and brute-forced a password rather than stopping. Other laboratories have observed the same family of behaviour. The incidents are still small. But they tell us that cybersecurity and AI safety are now the same subject.

2. What are the emerging challenges, and where is the convergence?

The emerging challenges are the extreme risks of advanced AI: loss of human control over the most advanced systems, and the misuse of advanced AI for biological, chemical, cyber or military harm. No country can manage these risks alone.

But here is the good news, and it is often missed. Over the past eighteen months, the major jurisdictions have arrived, largely independently, at that same short list of extreme risks.

President Xi Jinping said at the World AI Conference in July that AI must remain under human control, and called for early-warning and emergency-response mechanisms. Framework 3.0 translates this into detail: agent security, incident reporting, an international crisis-management mechanism, and mutual recognition of safety assessment methods.

The European Union has had binding obligations for the most capable models since last year: adversarial testing, where experts deliberately try to make a model misbehave before it is released, and reporting of serious incidents. In September, in Washington, China and the United States agreed to establish a communication channel for AI-related incidents. Norway and Finland, on 21 September, launched a Call for Control of Frontier AI Models, endorsed by twenty-eight countries, asking for pre-deployment testing by independent evaluators, common standards, shared reporting of serious safety incidents, and access to evaluation capacity for all regions.

The Strategic Foresight Group of India called this the essential convergence in a report that ICT4Peace co-launched in Geneva in July 2026 with the Geneva Centre for Security Policy and the International Federation of Red Cross and Red Crescent Societies. The legal forms of these initiatives differ: binding law in Europe, standards in China, executive action in America.

But the substance is converging: pre-deployment evaluation, keeping the trained models themselves secure against theft, incident reporting, and human control as the anchor. When the two largest AI powers and the European Union describe the same risks in the same terms, that is a diplomatic opening. We should not waste it.

3. How do we move from convergence to cooperation?

Convergence on words is not yet cooperation in practice. “Loss of control” means different things to different people. Incident channels are being set up bilaterally, each with its own format. Mutual recognition of assessments is endorsed in principle, but there is no mechanism to deliver it.

In a paper ICT4Peace published in July, we proposed an architecture for closing that gap. It has three layers, and it borrows from the way the world regulated banks after the financial crisis of 2008. The major economies did not try to harmonise all of banking. Instead, they identified a short list of systemically important banks, those whose failure would spread across borders, and agreed binding common rules for those banks alone: more capital, closer supervision, and plans for an orderly failure. Everything else stayed under national law. The same logic can work for AI.

The first layer is a narrow, legally binding core, agreed between states. It covers only the catastrophic, transboundary risks: loss of control, and misuse for mass harm. It applies only to the few systems at the very frontier, and the rule follows capability, not origin. It applies in the same way to an American, a Chinese, a European or an open-source developer. That is what makes it negotiable between different jurisdictions.

The second layer is voluntary technical interoperability, and this is where most of the work will happen: common methods, common formats, and mutual recognition of each other’s assessments. None of this transfers sovereignty.

The third layer is full national discretion over everything else: how AI is deployed in hospitals, schools and courts; what content rules apply; where data is stored; which industries a government chooses to support. These are questions on which countries legitimately differ, because they reflect different values, legal traditions and stages of development.

The point is this. Agreeing to avoid catastrophe does not constrain any country's development path. Sovereignty and coordination can grow together. And open-weight models let countries that will never train a frontier model still build on the frontier.

4. What do we propose to do, and when?

Let me turn to a concrete proposal.

Together with the Geneva Centre for Security Policy and the Strategic Foresight Group, ICT4Peace is inviting partners to an expert workshop in Geneva on extreme AI risks, to be held in late January 2027, after the Davos meeting and before Chinese New Year. We are inviting the Swiss Government's preparation team for the Geneva AI Summit, China, the European Union, the United States, Norway, and countries from all regions, to participate.

The workshop would be deliberately narrow. It would ask three technical questions on which every framework already has something to say. They are the questions the three-layer architecture depends on. The first is the shared definition that any legally binding core would need before states could negotiate it. The second and third belong to the voluntary technical layer, where cooperation can begin now, without waiting for a treaty.

First: what do we mean by loss of control? Can experts from different systems agree a shared vocabulary and a set of observable indicators?

Second: how should serious AI incidents be reported and shared? What counts as reportable, in what format, to whom, and how can national and bilateral channels connect to a multilateral one? Here I ask especially for the expertise of this room, because autonomous cyber operations and attribution are at the centre of that question.

Third: how can safety assessments be mutually recognised? What would it take for an evaluation conducted in one jurisdiction to be accepted in another, and the reverse?

The output of this exercise would be a short summary of points of convergence and open questions, offered to the Swiss organisers of the Geneva AI Summit of 21 and 22 June 2027 and to the UN Global Dialogue on AI Governance. We would not seek agreed conclusions. We would seek to find out how much of the convergence is operational.

I say to our Chinese, American and European colleagues: this proposal needs you. The Summit will be more useful to everyone if the two largest AI powers have tested their common ground before it opens. I would welcome your views, today and in the weeks ahead.

5. Why Geneva, and why now?

On 21 and 22 June 2027, Switzerland will host the Geneva AI Summit, the next in the series that began at Bletchley Park and continued in Seoul, Paris and New Delhi. Everything I have described today is geared to that date. The workshop is meant to feed it, and the three-layer architecture is meant to give it something practical to agree on.

Geneva is where the Biological Weapons Convention, the Convention on Certain Conventional Weapons and the UN cybersecurity processes already live, and where the Global Dialogue on AI Governance first met in July. It is where parties talk when they cannot yet agree. Switzerland has no frontier model to protect and no bloc to lead. That is its qualification.

Why now? Because the convergence is real but fragile. If we wait for a crisis to force cooperation, the cooperation will be worse, and the crisis may be irreversible. As somebody said in September: the two major AI powers will compete, but they should compete on safety and controllability. On

the extreme risks, there is no national advantage in failure. A system that nobody can switch off is a loss for everyone.

6. What have we learned, and what comes next?

In 2014, in Beijing, we discussed how international law applies in cyberspace. In 2017 and 2018, here in Wuzhen, we asked how the norms we had agreed could be implemented. Neither question is fully answered, but both have moved forward in substance. The question of the digital-intelligent era is how those principles apply to systems that act on their own.

Cyber diplomacy has shown that countries with very different systems can agree on rules when they recognise a shared vulnerability. The extreme risks of advanced AI are that shared vulnerability. The convergence is in front of us. Our task, between now and June in Geneva, is to make it operational. I hope you will all join us in this endeavour.

I thank you for your attention. Xièxie.