WSIS FORUM 2026 | 6-10 July 2026 Geneva, Switzerland

Session Outcome Document

Who Certifies the AI That Moves Among Us? Building Independent Safety Assurance for AI in the Physical World

Responsible AI for Safety and Ethics (RAISE) – Eric Stefanello
Thursday, 9 July 2026, 13:00–13:45 (UTC+02:00) — Room K, Palexpo — Session 208
https://www.itu.int/net4/wsis/forum/2026/Agenda/Session/208

Key Issues Discussed

- AI has left the screen: autonomous and assistive systems now make consequential decisions in the physical world — drones over inhabited areas, ground robots in hospitals, airports and streets, surgical and assistive devices — often around civilians who never consented to their presence.
- A structural gap exists in the global governance architecture: aviation, nuclear power and rail each rest on a mature, independent safety-certification regime, yet no equivalent body certifies AI systems acting in open physical environments. The EU AI Act addresses governance and risk classification; ISO/IEC standards are emerging; EASA has begun work for aviation — but the general cyber-physical case falls between existing mandates.
- WSIS Action Line C5 ("confidence and security in the use of ICTs") has meant cybersecurity for two decades; it must now extend to safety in the engineering sense — the assurance that an autonomous system will not harm those around it.
- Safety is not a property of the model alone: contemporary AI models are not certifiable the way classical critical software is, because of the gap between the mathematics that builds a system and the mathematics that would guarantee its behaviour at scale. Safety is therefore a property of the integrated architecture and its verified operating envelope.
- The evaluation methods already exist — EASA's SORA methodology shows how to assess the operational risk an autonomous system poses to people. What is missing is not the science but the institutional home to apply it independently to the general case.
- The multilateral system established two AI mechanisms in 2025–2026 — the Independent International Scientific Panel on AI and the Global Dialogue on AI Governance — but neither produces, nor is mandated to produce, an operational safety guarantee for a given system. A third pillar — independent third-party certification — remains empty.

Key Outcomes of the Session

- *Key insight:* independence is a structural necessity, not a virtue added afterwards. The maritime industry separated insurer from certifier in 1834 — one cannot impartially judge the safety of what one has an interest in deploying. This principle later built modern aviation and safety-critical software certification, and applies directly to AI: the certifier cannot be the developer, its insurer, or its client.

- *Key insight:* certification must be graduated and pre-normative — calibrated to operational risk (in the spirit of SORA) and producing safety criteria upstream, before standards harden, so that future norms rest on a scientifically grounded basis rather than a consensus negotiated under pressure.
- *Key insight:* credible certification enables deployment rather than obstructing it, exactly as aviation certification made mass air travel possible by making it trustworthy.
- *New initiative presented:* RAISE (Responsible AI for Safety and Ethics), a Swiss foundation forming in Geneva, was presented as an effort to instantiate this missing certification pillar — conducting the pre-normative work of defining safety criteria for AI interacting with the physical world, drawing on the 1834 separation principle and SORA-type operational risk assessment.
- *Commitment:* RAISE positions itself as a concrete contribution to implementing the AI dimension of the WSIS+20 review, and to anchoring an independent safety-assurance capability within International Geneva.

Key Recommendations and Forward-Looking Actions

- Recognise safety assurance for physically-acting AI as the unaddressed half of WSIS Action Line C5, and name it explicitly in the WSIS+20 follow-up toward 2035.
- Mandate and recognise an independent, third-party safety-certification function for cyber-physical AI — complementing, not duplicating, the Scientific Panel (which produces knowledge) and the Global Dialogue (which produces deliberation).
- Articulate this pre-normative work with existing standards and regulatory frameworks (ISO/IEC, EASA, the AI Act), through technical dialogue among peers rather than institutional merger.
- Support the prefiguration of an independent certification institution based in Geneva, at the heart of the multilateral system, as a tangible demonstration of multilateralism's capacity to produce concrete public goods.
- Move from research to implementation: treat independent safety certification as operational infrastructure to be built now, before autonomous systems scale further into shared civilian space.

Reference Note — A Structural Gap in the Global Digital Architecture

The following develops the session's argument for readers who did not attend, and is intended to be citable in its own right.

Artificial intelligence is no longer confined to screens and data centres. It now flies drones over inhabited areas, drives ground robots through hospitals, airports and city streets, and operates assistive and surgical systems — taking decisions that act directly on the physical world, often in the presence of civilians who never consented to be near them. Aviation, nuclear power and rail each rest on a mature, independent safety-certification regime. For AI acting in open physical environments, no equivalent body exists. The European AI Act addresses governance and risk classification; ISO/IEC standards are emerging; EASA has begun work for aviation. The general case — an AI making consequential physical decisions among people — falls between existing mandates. This is not a transient regulatory lag; it is a structural gap in a governance architecture the international community is actively building.

The difficulty runs deeper than institutional design. A contemporary AI model cannot be certified the way a classical critical-software component can. There is a fundamental asymmetry between the constructive mathematics that builds and trains a system and the analytical mathematics that would allow its behaviour to be guaranteed at scale. This analytical deficit is the root reason such models are not certifiable in isolation. Safety, therefore, is not a property of the model: it is a property of the integrated architecture and its verified operating envelope. For learning systems, the objective itself cannot be fully specified without drift — Goodhart's law here is demonstrated rather than metaphorical — but verifiable bounds on

action can be imposed. The object to be certified shifts accordingly: not the learned policy, but the bounds that contain it.

Crucially, the methods to do this already exist. EASA's SORA methodology shows how to assess the operational risk a drone poses to people on the ground. What is missing is not the science but the institutional home to apply it, systematically and independently, to the general cyber-physical case. The multilateral system has recently equipped itself with two valuable mechanisms — the Independent International Scientific Panel on AI, which synthesises scientific knowledge on a non-prescriptive basis, and the Global Dialogue on AI Governance, which convenes multistakeholder deliberation. Neither produces, nor is meant to produce, an operational guarantee that a particular system is safe for the people it acts upon. That is a third and distinct pillar — independent third-party certification — and it remains empty. The distinction is exact: a scientific panel characterises a class of risks, as the IPCC characterises seismic hazard; it does not certify that a given structure will stand.

The shape of the solution can be deduced rather than invented. Independence is structural: the maritime industry separated insurer from certifier in 1834, recognising that one cannot impartially judge the safety of what one has an interest in deploying — the principle that later underpinned aviation and safety-critical software. The certifier of an AI system can therefore be neither its developer, nor its insurer, nor its client. Certification must be graduated to operational risk and pre-normative in character, supplying the technical substance that standards bodies and regulators will need. And far from obstructing innovation, credible certification is what permits deployment at scale — as aviation certification did not constrain air travel but made its mass adoption possible by making it trustworthy.

Such an institution belongs at the heart of the multilateral system, not within a national jurisdiction that would compromise its perceived independence. Geneva — home to the ITU, with its tradition of neutral standardisation and its unique density of international expertise — offers exactly these conditions. It is in this spirit that RAISE (Responsible AI for Safety and Ethics), a Swiss foundation, is forming in Geneva: to carry out the pre-normative work of defining safety criteria for real-world AI and to instantiate the missing certification pillar. The call that follows is deliberately selective rather than universal. It asks states and the multilateral system to recognise and mandate an independent certification function; it asks standards bodies and regulators to articulate their frameworks with this pre-normative work; and it asks funders to make its prefiguration possible. As the WSIS+20 review moves from stocktaking to implementation, safety assurance for AI acting in the physical world is the frontier on which "confidence and security" must now be earned.