

Internationale KI-Sicherheit mit Schweizer Führungsanspruch gestalten

Text: Der Bundesrat wird beauftragt, am KI-Gipfel 2027 in Genf aktiv auf international abgestimmte Mindeststandards für die extremen Risiken der leistungsfähigsten KI-Modelle hinzuwirken, die sich an deren Fähigkeiten und Auswirkungen orientieren. Er unterstützt die Weiterentwicklung eines internationalen Netzwerks, das geteilte Sicherheitsstandards sowie unabhängige Evaluation und Verifikation sicherstellt. Bestehende Schweizer Initiativen für Rechenkapazität, vertrauenswürdige Daten, offene Modelle und Kompetenzaufbau stärkt er als ergänzende technische Infrastruktur, organisatorisch getrennt von Evaluation und Verifikation. Er stellt die erforderlichen politischen, personellen und finanziellen Ressourcen sicher, damit die Schweiz diesen Prozess mit Ambition mitgestalten und vorantreiben kann.

Begründung: Um sein Gipfelziel einer «vertrauenswürdigen und verantwortungsvollen Entwicklung und Anwendung von KI» zu erreichen, soll der Bundesrat bei Mindeststandards für die extremen Risiken der Frontier-Modelle ansetzen: Missbrauch für Bio- und Cyberwaffen, Kontrollverlust, militärische Autonomie sowie potenziell weitreichende gesellschaftliche und wirtschaftliche Auswirkungen. Ein tragfähiger Ansatz orientiert sich an überprüfbaren Aussagen über ein System statt am Modell als Ganzes.

Organisatorisch sind Sicherheit, Evaluation und Verifikation zu trennen von geteilter Rechenleistung und offenen Modellen, etwa ICAIN und Apertus. Diese sichern Ländern ohne eigene Modelle Zugang und Kompetenzaufbau und wirken Machtkonzentration entgegen. Netzwerk heisst: Die Prüfkapazitäten bleiben in den beteiligten Staaten, die Standards werden gemeinsam gesetzt.

Die Schweiz hat mehrfach gezeigt, dass sie internationale Regeln für Hochrisikobereiche ermöglichen kann: als Verhandlungsort für Nuklear- und Chemiewaffenabkommen und als Sitz des Basler Ausschusses für Bankenaufsicht, der für systemrelevante Banken verschärfte Standards setzt. Auf KI übertragen heisst das: strengere Anforderungen für die grössten Risiken, unternehmerische Freiräume für alle übrigen Anwendungen. Als Sitz eines glaubwürdigen, unabhängigen Netzwerks würde die Schweiz auch als Innovationsstandort profitieren. Eine international abgestimmte Regelung könnte zudem einseitigen und diskriminierenden Zugangsbeschränkungen einzelner Staaten vorbeugen und verhindern, dass Anbieter Staaten gegeneinander ausspielen.

Traduction non officielle générée par l'IA

Façonner la sécurité internationale de l'IA avec une ambition de leadership suisse

Texte : Le Conseil fédéral est chargé d'œuvrer activement, lors du Sommet sur l'IA 2027 à Genève, à l'adoption de normes minimales harmonisées au niveau international pour les risques extrêmes liés aux modèles d'IA les plus performants, normes qui se fondent sur les capacités et les effets de ces modèles. Il soutient le développement d'un réseau international garantissant des normes de sécurité communes ainsi qu'une évaluation et une vérification indépendantes. Il renforce les initiatives suisses existantes en matière de capacités de calcul, de données fiables, de modèles ouverts et de développement des compétences en tant qu'infrastructure technique complémentaire, séparée sur le plan organisationnel de l'évaluation et de la vérification. Il met à disposition les ressources politiques, humaines et financières nécessaires pour que la Suisse puisse façonner et faire avancer ce processus avec ambition.

Développement : Pour atteindre l'objectif qu'il a fixé pour le sommet, à savoir un « développement et une utilisation de l'IA dignes de confiance et responsables », le Conseil fédéral doit se concentrer sur des normes minimales pour les risques extrêmes des modèles de pointe (frontier models) : utilisation abusive à des fins d'armes biologiques et cybernétiques, perte de contrôle, autonomie militaire ainsi que des effets sociétaux et économiques potentiellement considérables. Une approche viable se fonde sur des affirmations vérifiables concernant un système plutôt que sur le modèle dans son ensemble.

Sur le plan organisationnel, la sécurité, l'évaluation et la vérification doivent être séparées de la puissance de calcul partagée et des modèles ouverts, tels qu'ICAIN et Apertus. Ces derniers garantissent aux pays qui ne disposent pas de leurs propres modèles l'accès à ces technologies et le développement des compétences, et contrecarrent la concentration du pouvoir. Parler de réseau signifie que les capacités de contrôle restent dans les États participants, tandis que les normes sont fixées en commun.

La Suisse a démontré à plusieurs reprises qu'elle est en mesure de rendre possibles des règles internationales dans des domaines à haut risque : en tant que lieu de négociation des accords sur les armes nucléaires et chimiques, et en tant que siège du Comité de Bâle sur le contrôle bancaire, qui fixe des normes renforcées pour les banques d'importance systémique. Transposé à l'IA, cela signifie des exigences plus strictes pour les risques les plus importants et une marge de manœuvre entrepreneuriale pour toutes les autres applications. En accueillant un réseau crédible et indépendant, la Suisse en tirerait également profit en tant que pôle d'innovation. Une réglementation harmonisée au niveau international pourrait en outre prévenir les restrictions d'accès unilatérales et discriminatoires imposées par certains États et empêcher que des fournisseurs ne montent les États les uns contre les autres.