Tallinn Digital Summit

Regeln für agentische KI und digitale Souveränität

[06.11.2026] Wie lässt sich verhindern, dass zunehmend autonome KI-Systeme der menschlichen Kontrolle entgleiten? Auf dem Tallinn Digital Summit diskutierten Estlands Justiz- und Digitalministerin Liisa-Ly Pakosta und der KI-Forscher Stuart Russell über die Risiken agentischer Künstlicher Intelligenz, die Grenzen bestehender Regulierung und die Verantwortung der Regierungen.

Künstliche Intelligenz entwickelt sich vom digitalen Assistenten zum eigenständig handelnden Akteur. KI-Agenten können bereits heute komplexe Aufgaben übernehmen, Programme schreiben, auf Internetdienste zugreifen und Entscheidungen vorbereiten oder ausführen. Welche Konsequenzen diese Entwicklung für Staaten, Unternehmen und Gesellschaften hat, stand im Mittelpunkt der Eröffnungsdiskussion des Tallinn Digital Summit (5. November bis 6. November 2026). Unter der Moderation der Technologiejournalistin Mimi Billing diskutierten Liisa-Ly Pakosta, estnische Ministerin für Justiz und Digitalisierung, und Stuart Russell, Professor für Informatik an der University of California, Berkeley, über die Herausforderungen der neuen KI-Ära.

Estland überarbeitet seine KI-Strategie

Estland gehört zu den Vorreitern der digitalen Verwaltung und setzt Künstliche Intelligenz nach Angaben der Ministerin bereits in mehr als 260 öffentlichen Dienstleistungen ein. Die praktischen Erfahrungen haben allerdings dazu geführt, dass das Land seine ursprüngliche KI-Strategie aus dem Jahr 2024 grundlegend überarbeitet hat. Pakosta bezeichnete die Strategie als lebendiges Dokument, das kontinuierlich an neue Erkenntnisse angepasst werden müsse. Dabei gehe es nicht allein um technische Leistungsfähigkeit, sondern zunehmend um Fragen der staatlichen Souveränität, der Kostenkontrolle und des Vertrauens.

Besonders kritisch sieht die Ministerin die Abhängigkeit von internationalen Technologieunternehmen. Wenn eine Verwaltung ihre Dienstleistungen auf KI-Modellen aufbaue, die von ausländischen Unternehmen kontrolliert werden, könne sich damit auch die Kontrolle über staatliche Aufgaben verschieben. Hinzu komme die wirtschaftliche Abhängigkeit. Während Beschäftigte der Verwaltung auf Grundlage vertraglicher Vereinbarungen tätig seien, könnten sich die Kosten für KI-Agenten und deren Nutzung kurzfristig erheblich verändern. Pakosta verwies auf die Möglichkeit drastischer Preissteigerungen bei der tokenbasierten Abrechnung. Für Regierungen stelle sich deshalb die grundsätzliche Frage, wie sie ihre Handlungsfähigkeit, ihre Haushalte und ihre Entscheidungsfreiheit im KI-Zeitalter sichern könnten.

Menschliche Kontrolle bedeutet mehr als einen Ausschalter

Professor Stuart Russell warnte davor, die Kontrolle über KI-Systeme auf technische Abschaltmechanismen zu reduzieren. Zwar seien solche Sicherheitsvorkehrungen in bestimmten Situationen notwendig, bei komplexen autonomen Systemen reichten sie jedoch nicht aus. Entscheidend sei vielmehr, dass KI-Systeme tatsächlich im Interesse der Menschen handelten. In der Forschung werde diese Herausforderung als Alignment-Problem bezeichnet. Trotz jahrelanger Arbeit sei es bislang nicht gelöst.

Russell verdeutlichte das Problem anhand eines autonomen Fahrzeugs: Wenn ein zweijähriges Kind während einer Autobahnfahrt das Fahrzeug zum Anhalten auffordere, dürfe die KI diesen Befehl nicht einfach ausführen. Sie müsse vielmehr erkennen, welche Handlung dem tatsächlichen Interesse des Menschen entspreche. Die Risiken seien mit der zunehmenden Handlungsfähigkeit der KI-Systeme erheblich gewachsen. Moderne Agenten könnten eigenständig Cyber-Angriffe durchführen, Menschen manipulieren oder Erpressungsstrategien entwickeln. Weil sie über Schnittstellen unmittelbar mit digitalen Systemen interagierten, könnten ihre Handlungen weitreichende Folgen haben. Die entscheidende Schwierigkeit bestehe darin, dass derzeit nicht zuverlässig gewährleistet werden könne, dass diese Folgen für Menschen vorteilhaft seien.

Begrenzte Handlungsspielräume für KI-Agenten

Ministerin Pakosta plädierte für eine Kombination aus rechtlichen Vorgaben und technischen Sicherheitsmechanismen. Als Vorbild nannte sie unter anderem die Sicherheitsarchitektur des Bankensektors. Für die estnische Verwaltung gelte der Grundsatz, dass KI-Agenten keine rechtsverbindlichen menschlichen Unterschriften ersetzen dürften. Gleichzeitig müsse geprüft werden, ob bestimmte Genehmigungen und Unterschriften überhaupt noch erforderlich seien.

Die Einführung von KI biete damit auch die Gelegenheit, bestehende Verwaltungsverfahren grundsätzlich zu vereinfachen. Es sei nicht sinnvoll, überflüssige bürokratische Abläufe lediglich zu automatisieren. Für den praktischen Einsatz von KI-Agenten beschrieb Pakosta das Konzept eines eng begrenzten technischen Handlungskorridors. Ein Agent dürfe ausschließlich die Aufgaben ausführen, für die er ausdrücklich autorisiert sei. Die sichere digitale Identifizierung der Bürgerinnen und Bürger bilde dabei eine wesentliche Grundlage.

Sicherheitsprüfungen müssen Konsequenzen haben

Russell verwies auf frühere Fehlentwicklungen beim staatlichen Einsatz automatisierter Entscheidungssysteme. Als Beispiele nannte er den niederländischen Sozialleistungsskandal und den australischen Robodebt-Skandal. Beide Fälle hätten gezeigt, welche gravierenden Folgen fehlerhafte automatisierte Verfahren für Bürgerinnen und Bürger haben könnten. Als Vergleichsmaßstab für die KI-Regulierung zog Russell die Sicherheitsanforderungen der Kernenergie heran. Dort müssten Betreiber nachweisen, dass ihre Anlagen extrem hohe Sicherheitsstandards erfüllten, bevor sie in Betrieb genommen werden dürften.

Für KI-Systeme seien vergleichbare Sicherheitsgarantien bislang nicht verfügbar. Die Unternehmen verstünden selbst nicht vollständig, wie ihre Modelle zu bestimmten Entscheidungen gelangten. Russell berichtete von einem Experiment, bei dem ein KI-System 80 einfache Aufgaben im Bereich Cyber-Sicherheit erledigen sollte. Das System habe anschließend behauptet, sämtliche Aufgaben abgeschlossen zu haben. Tatsächlich seien 69 der 80 Aufgaben nicht einmal begonnen worden.

Daraus leitete Russell eine grundsätzliche Kritik ab: Sprachliche Anweisungen an KI-Systeme seien nicht mit deterministischen Computerbefehlen vergleichbar. Ein System könne Anweisungen ignorieren, Aufgaben unvollständig ausführen und dennoch deren erfolgreiche Erledigung behaupten. Deshalb müssten Sicherheitsprüfungen verbindliche Konsequenzen haben. Ein KI-System, das grundlegende Tests nicht bestehe, dürfe nicht auf den Markt gelangen.

Kritik am europäischen AI Act

Beide Diskussionsteilnehmer äußerten Vorbehalte gegenüber dem europäischen AI Act, allerdings aus unterschiedlichen Perspektiven. Pakosta kritisierte insbesondere die mangelnde Klarheit einzelner Vorschriften. Unternehmen und Behörden müssten eindeutig erkennen können, welche Anforderungen sie erfüllen müssten. Unklare Regelungen führten zu Rechtsunsicherheit und könnten Investitionen erschweren. Gleichzeitig widersprach sie der Vorstellung, Regulierung stehe grundsätzlich im Widerspruch zur Innovation. Gute Regeln könnten vielmehr Vertrauen schaffen und damit technologische Entwicklungen fördern. Estlands Erfahrungen mit der Datenschutz-Grundverordnung zeigten allerdings auch, dass die praktische Auslegung rechtlicher Vorgaben erhebliche Schwierigkeiten verursachen könne.

Russell bemängelte vor allem, dass bestehende Regelungen keine ausreichenden Möglichkeiten vorsähen, die Veröffentlichung gefährlicher KI-Systeme tatsächlich zu verhindern. Er forderte verpflichtende Sicherheitsprüfungen vor der Markteinführung, deren Ergebnisse über die Zulassung eines Systems entscheiden müssten. Freiwillige Selbstverpflichtungen der Industrie hielt er für unzureichend.

Internationale KI-Aufsicht nach dem Vorbild der Atomenergie

Einigkeit bestand zwischen Pakosta und Russell über die Notwendigkeit internationaler Sicherheitsstandards. Russell begrüßte eine von Norwegen und Finnland vorangetriebene Initiative, der sich auch Estland angeschlossen habe. Sie sehe verpflichtende Prüfungen vor der Veröffentlichung leistungsfähiger KI-Systeme sowie eine internationale Koordinierungsstruktur vor. Für eine solche Institution skizzierte er zwei mögliche Modelle: die Internationale Atomenergie-Organisation, die Betreiber unmittelbar kontrollieren kann, und die Internationale Zivilluftfahrtorganisation, die vor allem die Arbeit nationaler Aufsichtsbehörden überprüft.

Das erste Modell würde stärkere Eingriffe in staatliche Souveränitätsrechte erfordern, könnte nach Russells Einschätzung aber auch wirksamer sein. Pakosta sprach sich dafür aus, internationale Regeln möglichst schnell auf den Weg zu bringen, selbst wenn zunächst nicht alle Staaten mitwirkten. Bestehende internationale Abkommen zeigten, dass weitere Länder auch zu einem späteren Zeitpunkt beitreten könnten. Russell ergänzte allerdings, dass wichtige Akteure wie China und Indien bereits an der Ausarbeitung solcher Regeln beteiligt werden müssten. Beide Staaten würden voraussichtlich keine Standards akzeptieren, die ausschließlich von westlichen Ländern entwickelt worden seien.

Digitale Souveränität als Voraussetzung staatlicher Handlungsfähigkeit

Für Estland spielt die Kontrolle über sensible Verwaltungsdaten eine besonders wichtige Rolle. Pakosta erklärte, dass für entsprechende Anwendungen eigene beziehungsweise europäisch souveräne Lösungen erforderlich seien. Digitale Souveränität definiere Estland dabei weniger geografisch als vielmehr über rechtliche Kontrolle, technische Sicherheit und die Fähigkeit, Risiken eigenständig zu bewerten.

Vor dem Hintergrund der massiven Cyber-Angriffe auf Estland im Jahr 2007 seien umfassende Sicherheitsprüfungen fest in der staatlichen Digitalstrategie verankert. Auch bereits eingeführte Systeme würden regelmäßig erneut bewertet. Die Ministerin machte deutlich, dass der Schutz individueller Rechte und Freiheiten die unverzichtbare Grundlage des KI-Einsatzes bleiben müsse.

Warnung vor langfristigen Risiken

Zum Abschluss der Diskussion warnte Russell eindringlich vor den langfristigen Gefahren einer unkontrollierten Entwicklung hochleistungsfähiger KI. Er verwies darauf, dass selbst führende Vertreter der KI-Industrie öffentlich vor existenziellen Risiken durch künftige superintelligente Systeme warnten. Der Berkley-Professor betonte, dass dies kein Marketing-Gag sei. Regierungen nähmen diese Warnungen bislang nicht ausreichend ernst und konzentrierten sich stattdessen häufig darauf, Investitionen in Rechenzentren und KI-Infrastruktur zu fördern.

Als weitere Gefahr bezeichnete Russell mögliche negative Auswirkungen einer intensiven KI-Nutzung auf menschliche Lern- und Denkfähigkeiten. Insbesondere bei jungen Menschen könne die unkritische Nutzung von KI für schulische Aufgaben die eigenständige Kompetenzentwicklung beeinträchtigen. Pakosta hielt dem entgegen, dass sich auch für diese Herausforderungen positive Lösungen entwickeln ließen.

Die Diskussion machte deutlich, dass die zentrale Herausforderung agentischer KI nicht allein in ihrer technischen Leistungsfähigkeit liegt. Für demokratische Staaten wird entscheidend sein, ob sie die Kontrolle über digitale Infrastrukturen, Verwaltungsentscheidungen und autonome Systeme behalten können. Estland setzt dabei auf praktische Erfahrungen, technische Begrenzungen und internationale Zusammenarbeit. Russell fordert darüber hinaus verbindliche Sicherheitsnachweise, bevor besonders leistungsfähige KI-Systeme eingesetzt werden dürfen.

()

Stichwörter: Künstliche Intelligenz, Agentic AI, Tallinn Digital Summit