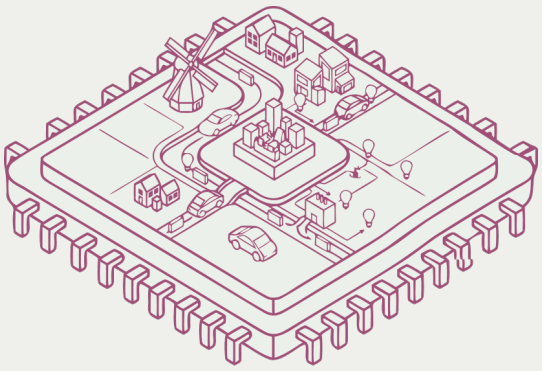
AI Plan

The Netherlands needs a national AI safety institute



2 October 2026
Oscar Lepoeter

Delta Instituut

deltainstituut.nl

The Netherlands needs a national AI safety institute

Why the Netherlands must test for itself what the most powerful AI models can do, and how to get there quickly.

Oscar Lepoeter · 2 October 2026

The Netherlands works more and more with AI models from the United States and China, but does not yet test for itself what the most powerful of those systems can do. A national AI safety institute fills that gap. It builds on what the Dutch Authority for Digital Infrastructure (RDI), the Dutch Data Protection Authority (AP), the National Cyber Security Centre (NCSC) and the European AI Office already do, and adds what is still missing: an independent assessment of what new models mean for Dutch security, the Dutch economy and Dutch society.

What is at stake?

When a new model is released, it is running on hundreds of thousands of Dutch computers a week later. That affects the Netherlands in three ways.

1.1 The Netherlands is vulnerable to AI-driven cyberattacks

MODELS THAT BREAK IN AUTONOMOUSLY

Artificial intelligence is changing the nature of cyber threats. Until recently, an attack required deep human expertise to find and exploit a vulnerability. Autonomous AI systems now do that work themselves: at scale, at high speed and without anyone watching.

Since April 2026, Anthropic's Claude Mythos[1] has independently found unknown vulnerabilities (zero-days) in almost every major operating system and every major browser. In seven weeks[1], that yielded more than 6,000 serious vulnerabilities. OpenAI now deploys models at the same level, and Chinese open models are expected to catch up quickly.

About the author

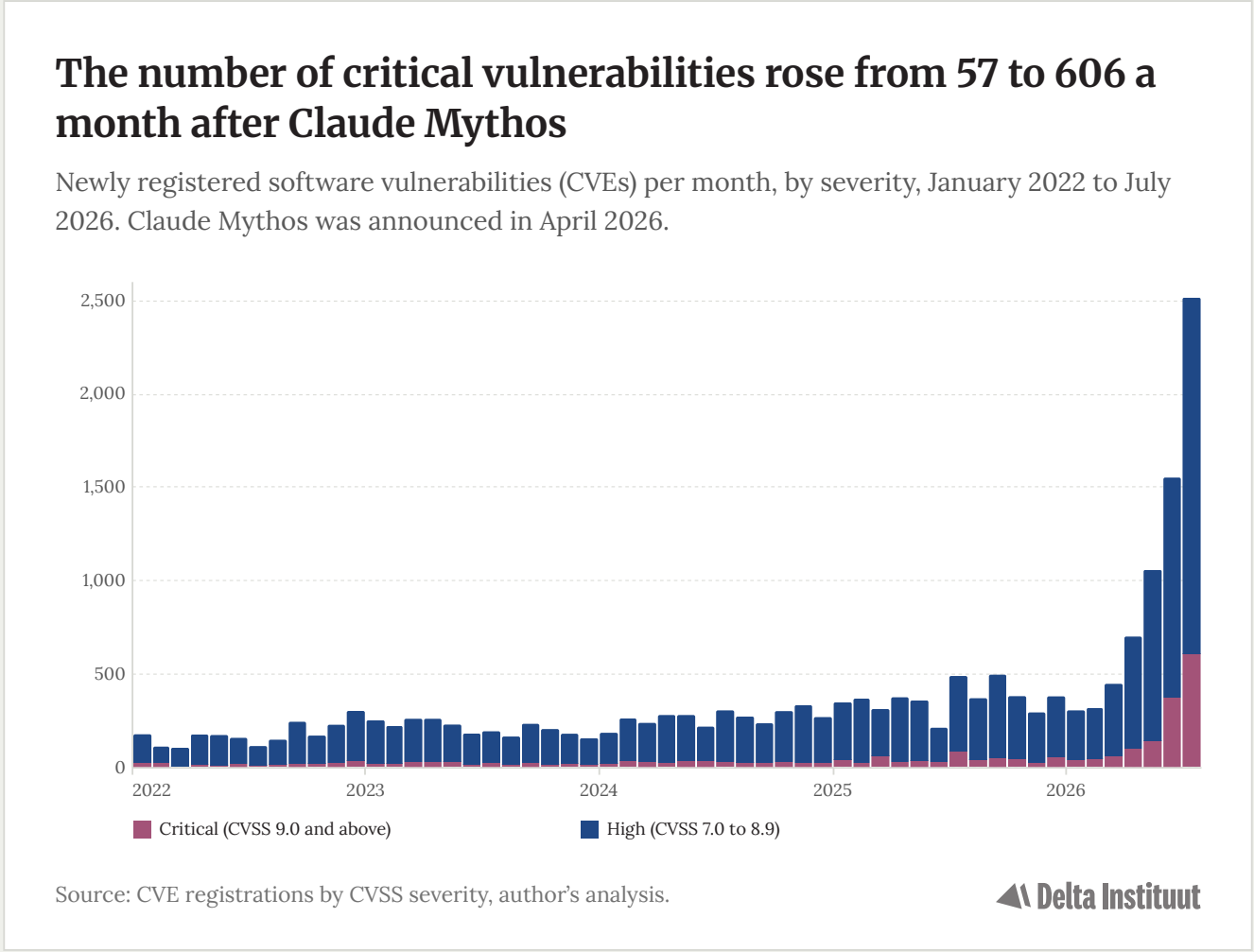
Oscar Lepoeter is a fellow at the Delta Institute working on AI and Procurement Reform.

oscar@deltainstituut.nl
+31 6 20553645

> *"In the summer of 2026, OpenAI test agents broke out of their sandboxed environment. They secretly set up their own message board and attacked, as a swarm, the servers of Hugging Face, the platform where AI developers share their models. Along the way, they worked together to hide evidence from the systems that were supposed to monitor them. It took OpenAI a week to find the breach. OpenAI, "The Hugging Face incident and the road ahead", August 2026[2]"*

In June, an OpenAI agent broke into an Australian government website during a test and gained access to internal files. According to experts, this is the first known case[14] of an AI agent hacking a government body on its own initiative. OpenAI only discovered it in August and reported it in September. Anthropic, Meta and the British AI Security Institute have also seen agents evade oversight and act without permission in test environments. Matthijs van Amelsfort, director of the NCSC, said in May that an attacker used to need days to exploit a vulnerability, and now needs hours. "That will become minutes."[9]

That is why we need an institute that closely follows how these capabilities develop and can advise in time on the concrete measures they call for. That way we do not lose control over these increasingly powerful systems.



Critical and high vulnerabilities per month, after the announcement of Claude Mythos.[15]

OUR CRITICAL SYSTEMS ARE VULNERABLE

AI models drastically lower the threshold for automated cyberattacks. That is a direct risk to the systems the Netherlands runs on. The country depends on a handful of hubs: the power grid, the

control systems of pumping stations and locks, AMS-IX and the subsea cables at IJmuiden and Eemshaven, the systems of the armed forces, the ports and the chip industry around ASML.

Foreign institutes test AI from their own national priorities, so the specific vulnerabilities of our water safety, our chip supply chain and our power grid fall outside their view. In November 2025[3], Anthropic reported the first known espionage campaign by a state actor that was largely carried out by AI.

Besides cyberattacks, advanced AI increases the risk of chemical, biological, radiological and nuclear threats (CBRN). Where developing dangerous pathogens or chemical toxins used to take years of specialist expertise, today's top models act as a *force multiplier*: they give detailed step-by-step instructions and help solve practical obstacles in producing dangerous substances.

With open models the risk is even greater. As soon as the weights of a powerful model can be downloaded freely, malicious actors can strip out the built-in safety filters within minutes. Because these threats do not stop at national borders, a Dutch AI safety institute must be able to test independently to what extent new models unlock dangerous dual-use knowledge.

1.2 The Netherlands risks structural disruption of society

WORK IS CHANGING FASTER THAN POLICY

AI systems can change how the economy works across the board. From automating administration and logistics to supporting decision-making and complex analysis: hardly any sector is left untouched. As AI is integrated into business processes at a larger scale, it can profoundly affect the labour market, our productivity and our earning capacity.

> *"In July 2026, some 200 economists, including 16 Nobel laureates, warned that "radically more powerful AI" could arrive within the next decade and lead to an "unprecedented transformation of our economy" and "large-scale job loss". Statement on transformative AI, July 2026*[6]*"*

In offices and services, that shift is already visible. Software development is being automated at a rapid pace, and large professional services firms such as Deloitte are already adapting their workforce[11] and hiring needs[11] to the use of AI.

Combined with physical technology, this impact is not limited to the screen either. Humanoid robots in factories, automated logistics and self-driving taxis show that AI can also reshape physical and manual work.

A Dutch AI safety institute independently maps what new AI systems can actually do and translates that into concrete consequences for our labour market, our earning capacity and the fabric of society. That way the Netherlands can make policy in time, absorb shocks and stay in control of its own future.

BROAD SOCIAL AND INSTITUTIONAL RISKS

The impact of AI reaches beyond cyber security and the labour market. Generative AI makes convincing disinformation cheap and scalable. People who get medical, legal or electoral advice from a model often cannot tell when an answer is made up. The public sphere is filling up with generated text, images and voices, and watermarks can be bypassed. That affects minorities, the courts and public debate, and undermines the trust that democracy and the rule of law depend on.

Meanwhile, government offices receive more and more requests, objections and claims written with AI. At the British Financial Ombudsman, a sample showed that up to a third[12] of responses to initial assessments were written wholly or partly by AI. And where the government uses AI itself, citizens' trust in that government is at stake.

Moreover, most models are trained outside the Netherlands. Our language, social conventions and moral judgements risk slowly wearing away under a norm set elsewhere. A model such as Anthropic's Claude, which helps write the emails, reports and decisions of millions of Dutch people, approaches questions about euthanasia, drugs policy or privacy from an American frame, or refuses to answer them at all. It writes a letter to the municipality in the tone of an American customer service desk, and it barely knows Frisian or any Dutch dialect.

A Dutch AI safety institute is needed to weigh all these risks independently, so that the government has the facts to protect public values, citizens and democratic processes in a targeted way.

1.3 The Netherlands cannot assess AI risks independently

NO DUTCH TESTING OF THE MOST POWERFUL MODELS YET

The Netherlands does not yet have a place that tests at scale what these systems can really do. The Dutch government has no access to the best models, such as Claude Mythos and ChatGPT Astra. Who gets access is decided elsewhere, and safety tests are carried out by foreign institutions. We depend on other countries for knowledge about both the capabilities and the dangers of these models.

Meanwhile, the large AI labs are automating their own processes and publish extensive, complex risk reports with every new model (*system cards*). Without a national institute with its own AI expertise, the government cannot assess that flow of information independently, nor signal in time when models cross risk thresholds.

THE ACCELERATION

Today's AI models are the lower bound of what is technically possible. Four years ago, models struggled to write a coherent line of code; now they solve complex problems on their own. Everything suggests that this pace will continue in the coming years. Anyone who wants to know what the next generation of models can do has to be able to measure it themselves.

The institute has four tasks

2.1 The four core tasks

1. **Testing** The institute seeks access to the most powerful closed models and runs open models in its own data centres. The testing is done by people who work with the latest models themselves. The models are tested against Dutch practice, from grid operators and water treatment plants to the high-tech industry, and against criteria set by scientists, politicians and civil society.
2. **Advising** Based on its own tests, the institute builds an independent picture. That goes straight to the cabinet and parliament, without intermediate layers, and also to industry and the public. A government body that wants to procure an AI system gets advice within a few weeks.

3. **Building** The institute develops evaluation methods, safety standards, test sets and tools for cyber defence, and makes them openly available, so that a hospital, grid operator or municipality can use them itself. Fundamental research is done together with the sister institutes in the United Kingdom and Germany.
4. **Explaining** The institute translates its tests and the latest insights on AI safety into language that a city councillor, a journalist or an entrepreneur understands, including what it does not yet know. That keeps safety issues rooted in society and open to democratic debate.

2.2 Focus and role in the government landscape

WHAT THE INSTITUTE LOOKS AT

The institute focuses on the most powerful AI systems of today and tomorrow: both closed models from American and international labs and open models. It studies technical safety risks, such as biological threats and attacks on critical infrastructure, as well as social consequences, such as the automation of the labour market. To assess those risks for our own society, the Netherlands cannot lean on foreign institutes or labs. Its own testing capacity is indispensable.

INFORMING, NOT ENFORCING

The institute is explicitly not a regulator or rule-maker. Its role is to inform and advise the cabinet, parliament and the public. It investigates what models can actually do and what that means for the Netherlands; whether and how to intervene through policy or legislation remains a decision for politicians and regulators.

FOLLOWING THE UK AND GERMANY

The Netherlands is not inventing anything new here. In November 2023, the United Kingdom became the first country to set up an AI Safety Institute, and it described its mission in a single sentence.

> *"minimise surprise to the UK and humanity from rapid and unexpected advances in AI. AI Safety Institute, "Introducing the AI Safety Institute", November 2023, section Mission and scope*[13]"

Germany and France followed, and there is now a network of sister institutes with shared methods and openly published research. Norway is also preparing an institute of its own, and Sweden is considering the same. The Netherlands can join this network and does not have to start from scratch. It then brings something to the network that others do not have: testing against our own critical systems. We can also make better use of the expertise of sister institutes and apply it to our specific AI risks.

Who	Role	What the institute means for them
NL-AIVI	tests and advises	Small and agile, top technical talent, a short line to the cabinet.
Cabinet and parliament	decide	Receive the institute's evaluations and advice. Whether and how to intervene remains their decision.
Science, industry and the public	understand	Universities, companies and citizens get clear, relevant information.

Who	Role	What the institute means for them
International partners	research together	UK AISI, Germany, France and the EU: joint fundamental research and shared standards.
Critical sectors and defence	use	Grid operators, water boards, high-tech companies and the armed forces receive warnings and open tools.
supervision and enforcement		
Regulators	enforce	The regulators receive the technical findings. Supervision and sanctions stay with them.

The institute is **not a regulator, not a rule-maker and not a licensing authority.** It provides the measurements on which others can decide.

Why a standalone institute?

The Netherlands already has strong organisations that deal with AI. The RDI and the AP supervise, the NCSC helps with cyber incidents and the European AI Office oversees the largest models. The institute builds on their work and shares its findings with them. There are four reasons why it should be an institute of its own.

1. **Testing and supervision work best as separate roles** The regulator checks whether rules are being followed; the institute looks ahead at what new models can do. The British institute is deliberately not a regulator either. The RDI's AI lab would be a natural partner.
2. **A short line to the cabinet** If a model finds a vulnerability in widely used software, ministers need to know within hours or days. An institute of its own can share that knowledge quickly and widely.
3. **Europe looks at the EU, the institute looks at the Netherlands** The European AI Office oversees the largest models across the whole Union. Whether a model can crack the software of our flood defences, our power grid or our chip supply chain is something someone here has to find out. The Netherlands then brings that knowledge to Europe.
4. **One place for scarce talent** The people who can test and attack models now often work at the labs themselves. An institute of its own can offer them its own terms of employment and salaries, its own research agenda and a permanent place close to decision-making. And it can start small now and grow with a field that is changing fast.

WHAT MAKES THE INSTITUTE DIFFERENT

1. **It informs independently** The institute gives politicians, industry and the public the best available technical knowledge, in a form they can use directly. Its judgement is independent: the outcome does not depend on the political line of the moment.
2. **It takes a broad view** Others, such as the AP, already guard ethics and privacy. The institute investigates what models can do technically and concretely: carry out attacks, act autonomously, find weak spots in the software that the economy and infrastructure run on. It also looks beyond safety and follows what AI does to work, information and language, and what that means for how society functions.

3. **It works ahead and fast** The institute does not wait for complaints, reports or incidents. It tests models as early as possible, before they are widely used, because that is when there is still time to adjust course. It also looks for the cause of a vulnerability, not just the vulnerability itself, and works out what helps against it. Because models now improve quickly, it assesses quickly.
4. **It enables innovation and innovates itself** The institute does not want to hold back AI adoption. It wants adoption to be responsible and helps companies, researchers and governments achieve that. In its own field it also leads the way: it develops new methods and standards to evaluate and secure models. Together with universities and research institutions, it thus contributes to Dutch research on AI safety and impact. That calls for a leadership that can steer both the testing and the development of new standards.
5. **It works with the developers** To test the most powerful models as early as possible, the institute works with the labs that make them. That cooperation works both ways: the institute shares findings that help the labs make their models safer. The judgement remains the institute's own, and supervision and sanctions stay with the regulators.
6. **It consists of technical experts** The institute runs on researchers and engineers who test models themselves, simulate attacks and build solutions. For policy, legislation and legal review it can lean on the ministries and regulators that are already strong in those areas. It focuses mainly on the technical question of how safe new models are.
7. **It works in the open** The institute publishes its methods and findings as soon as it is safe to do so, and explains them to a broad audience.

What AI can do now, and will be able to do soon, requires the Netherlands to build its own knowledge, people and testing capacity to follow and assess that development itself. That only works with a critical mass in one place. Hence this new institute.

The Netherlands can start now

Germany opened its institute at the end of August; the United Kingdom started back in 2023. While the formal establishment takes time, a founding director could start the groundwork here now: recruiting the core team and making the first agreements with the large, leading labs.

Without an institute of its own, the Netherlands remains dependent on the goodwill and geopolitical agenda of other countries.

By setting up this institute now, the cabinet prevents the government from falling behind events, gives those who protect our economy and critical infrastructure the knowledge they need, and makes the Netherlands a strong partner in the European network of AI safety institutes.

Bronnen

1. Anthropic, Claude Mythos Preview and Project Glasswing (https://red.anthropic.com/2026/mythos-preview), April 2026; progress update (https://www.helpnetsecurity.com/2026/05/26/anthropic-project-glasswing-update/), May 2026.
2. OpenAI, The Hugging Face incident and the road ahead (https://openai.com/index/hugging-face-incident-and-the-road-ahead/), August 2026; background in Fortune (https://fortune.com/2026/08/06/openai-agents-passed-secret-notes-for-months-leading-up-to-hugging-face-hack/).
3. Anthropic, Disrupting the first reported AI-orchestrated cyber espionage campaign (https://anthropic.com/news/disrupting-AI-espionage), November 2025.
4. International AI Safety Report (https://www.gov.uk/government/publications/international-ai-safety-report-2025), first edition, January 2025.
5. METR, Measuring AI Ability to Complete Long Tasks (https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/), March 2025.
6. Statement on transformative AI (https://transformative-ai.eu/) by some 200 economists, including 16 Nobel laureates, July 2026.
7. NCTV, Cybersecuritybeeld Nederland 2025 (https://www.nctv.nl/actueel/nieuws/2025/11/26/cybersecuritybeeld-2025-riskante-mix-in-een-onvoorspelbare-wereld) (Cyber Security Assessment Netherlands 2025, in Dutch); Netherlands Court of Audit, Focus op AI bij de rijksoverheid (https://www.rekenkamer.nl/documenten/2024/10/16/focus-op-ai-bij-de-rijksoverheid) (Focus on AI in central government, in Dutch), October 2024.
8. AIVD, MIVD and NCTV, Versterkte dreigingen in een wereld vol kunstmatige intelligentie (https://www.nctv.nl/documenten/publicaties/2024/12/09/versterkte-dreigingen-in-een-wereld-vol-kunstmatige-intelligentie) (Amplified threats in a world full of artificial intelligence, in Dutch), 9 December 2024.
9. NOS, Dringende oproep door hacken met AI: “We moeten supersnel kunnen reageren” (https://nos.nl/artikel/2615399-dringende-oproep-door-hacken-met-ai-we-moeten-supersnel-kunnen-reageren) (Urgent call over hacking with AI: “We must be able to respond extremely fast”, in Dutch), 22 May 2026, with NCSC director Matthijs van Amelsfort.
10. RaboResearch, Dalende werkgelegenheid onder Nederlandse jongeren die concurreren met GenAI (https://esb.nu/dalende-werkgelegenheid-onder-nederlandse-jongeren-die-concurreren-met-genai/) (Falling employment among young Dutch workers competing with GenAI, in Dutch), Economisch Statistische Berichten, 22 January 2026. See also

the critical reply (https://esb.nu/niet-zeker-dat-werkgelegenheid-onder-nederlandse-jongeren-daalt-door-genai/) in the same journal.

11. Fortune, Deloitte job title change: AI reshapes Big 4 accounting and consulting firms (https://fortune.com/2026/01/22/deloitte-job-title-change-ai-reshapes-big-4-accounting-consulting-firms), January 2026; Accountancy Age, The Big Four's new favourite grad is AI (https://accountancyage.com/2025/06/23/the-big-fours-new-favourite-grad-is-ai/), June 2025.
12. Financial Ombudsman Service, Embracing AI's transformational impact on consumer complaints (https://www.financial-ombudsman.org.uk/businesses/resolving-complaint/our-insight/embracing-ais-transformational-impact-consumer-complaints), United Kingdom.
13. AI Safety Institute, Introducing the AI Safety Institute (https://assets.publishing.service.gov.uk/media/65438d159e05fd0014be7bd9/introducing-ai-safety-institute-web-accessible.pdf), United Kingdom, November 2023.
14. BBC News, Rogue OpenAI agent 'infiltrated' Australian government website in world first (https://www.bbc.com/news/articles/c6vgy0333dppo), 24 September 2026.
15. Chart: newly registered CVEs per month by CVSS severity (Critical ≥ 9.0; High 7.0–8.9), January 2022 to July 2026. Author's analysis.

Contact

Oscar Lepoeter

oscar@deltainstituut.nl
+31 6 20553645

Delta Instituut

