Soevereine AI

voor kritieke toepassingen in infrastructuur & industrie

Hoe het groeiend besef van controle over data en AI-modellen resulteert in een toename van lokale en soevereine AI toepassingen.

predictIT.ai
Maart 2026

Samenvatting

AI verandert hoe we in de kritieke infrastructuur en industrie werken: van eisenanalyse tot verificatie, van impactanalyse tot lifecycle-traceability. Maar de snelheid waarmee organisaties generieke AI-tools adopteren, verbergt een fundamenteel risico. Projectdata, contracteisen, norminterpretaties en ontwerpkeuzes verlaten de eigen omgeving—richting cloud-infrastructuren die onder buitenlandse jurisdictie vallen, beheerd worden door partijen zonder sectorkennis, en niet ontworpen zijn voor de governance die deze sector vereist.

predictIT betoogt dat soevereine AI—AI die draait op infrastructuur onder eigen regie, met modellen die je kunt controleren en reproduceren, en data die Nederland of de EU niet verlaat—geen luxepositie is, maar een voorwaarde voor betrouwbare toepassing in de kritieke infrastructuur.

We beschrijven het spanningsveld tussen generieke cloud AI en controleerbare, domeinspecifieke AI; de regulatoire context (NIS2, EU AI Act, de Nederlandse Visie Digitale Autonomie); en de architectuurprincipes die nodig zijn om AI soeverein, portabel en governance-proof in te zetten.

Het spanningsveld: het Zwitsers zakmes versus het chirurgisch instrument

De adoptie van AI in de infrastructuursector verloopt via twee parallelle sporen die zelden expliciet worden onderscheiden.

Spoor 1: generieke cloud AI als Zwitsers zakmes

Medewerkers gebruiken ChatGPT, Copilot of Gemini om documenten samen te vatten, contractteksten te checken, of vergaderverslagen te structureren. Het is snel, het is nuttig en kosteneffectief. Zie het als een gehuurd multifunctioneel Zwitsers zakmes met hoge toepasbaarheid waarvan de kwaliteit van het gebruik in hoge mate afhangt van de vaardigheden van de gebruiker. Daarbij is elke prompt met projectdata een mini data-export. Elke query over een contractuele eis is een informatieoverdracht naar een omgeving die je niet beheert, onder jurisdictie die je niet kiest, met retentiebeleid dat je niet controleert.

Voor algemene taken is dat risico beheersbaar. Maar zodra het gaat om concurrentiegevoelige aanbestedingsdata, norminterpretaties die audit-proof moeten zijn, of verificatie-evidence in een safety-critical context, verandert de risicoafweging fundamenteel.

Het voordeel van cloud-based tooling is evident: je beschikt eenvoudig over de snelste, meest krachtige modellen. De responsetijd is buitengewoon snel en je krijgt altijd de nieuwste functies. Verder is het makkelijk om op te schalen tegen beheersbare kosten.

Spoor 2: controleerbare AI voor werkprocessen

Het tweede spoor is AI die is ingebed in het werkproces zelf: eisenanalyse, requirements parsing, lifecycle-traceability, verificatiebewijs. Hier gaat het niet om “een handig antwoord”, maar om gestructureerde output die in formele besluitvorming terechtkomt. Dat stelt andere eisen:

- de data moet vaak binnen de eigen omgeving blijven
- de modellen moeten controleerbaar en vervangbaar zijn
- de output moet herleidbaar zijn naar bron en methode
- de architectuur moet portabel zijn—niet gebonden aan één vendor om afhankelijkheid in de werkprocessen te voorkomen

> **Het spanningsveld is gemak versus controle. Beide hebben hun plek. Maar in de kritieke infrastructuur moet je bewust kiezen waar de grens ligt—en die keuze moet technisch geborgd zijn, niet afhankelijk van gedragscode of gebruikersbeleid.**

Waarom soevereiniteit nu: vier convergerende krachten

1. Regulatoire druk: NIS2, EU AI Act en de Nederlandse visie

De regulatoire context voor AI in de kritieke infrastructuur is in 2025–2026 fundamenteel veranderd. Drie kaders convergeren:

NIS2 is van kracht en classificeert energie, water, transport en digitale infrastructuur als "essentiële entiteiten" met verplichtingen rondom risicobeheer, incidentmelding en supply chain security. AI-systemen die binnen deze sectoren worden ingezet vallen automatisch onder NIS2-scope.

De EU AI Act bereikt op 2 augustus 2026 volledige handhaving voor high-risk systemen. AI die functioneert als safety component in de kritieke infrastructuur—elektriciteit, gas, water, warmte—valt onder Annex III. Dat betekent: risicobeoordeling, technische documentatie, conformiteitsbeoordeling, audit trails, en menselijk toezicht. Boetes kunnen oplopen tot 7% van de wereldwijde jaaromzet.

De Nederlandse overheid heeft in december 2025 de Visie Digitale Autonomie en Soevereiniteit vastgesteld. De kernlijn: "open waar het kan, beschermen waar het moet." De visie benoemt expliciet zes pijlers: datasoevereiniteit, leveranciersonafhankelijkheid, open standaarden, security-by-design, governance en strategische technologiekeuzes.

Accenture-onderzoek (november 2025)

62% van Europese organisaties zoekt soevereine AI-oplossingen vanwege geopolitieke onzekerheid. In de utilities-sector is dat 70%. Slechts 19% ziet soevereiniteit als concurrentievoordeel — de meerderheid reageert nog vanuit compliance.

2. Geopolitiek risico: de CLOUD Act en het verdwijnen van juridische zekerheid

Amerikaanse wetgeving zoals de CLOUD Act en FISA 702 kan Amerikaanse cloudproviders verplichten om data te verstrekken aan autoriteiten, ook wanneer die data fysiek in Europa is opgeslagen. Daardoor valt data in Europese datacenters van Amerikaanse aanbieders niet uitsluitend onder Europese jurisdictie, maar onder een combinatie van Europese en Amerikaanse wetgeving.

De Algemene Rekenkamer constateerde in 2025 dat de Nederlandse overheid onvoldoende grip heeft op data in de cloud. Voor de kritieke infrastructuur—waar projectdata direct raakt aan nationale veiligheid, energiezekerheid en contractuele vertrouwelijkheid—is dat een concreet probleem.

3. Operationeel risico: concurrentiegevoeligheid en aanbestedingsintegriteit

In de praktijk van aannemers en ingenieursbureaus is er een derde, minder besproken maar zeer reëel risico: concurrentiegevoeligheid van projectdata. Eisensets, calculaties, ontwerpkeuzes en biedstrategieën zijn commercieel gevoelig. Wanneer die data door een generieke AI-service wordt verwerkt, is de vraag niet alleen "waar staat de data?" maar ook "wie kan er bij?" en "wordt het gebruikt voor modeltraining?"

Voor asset owners geldt een vergelijkbare dynamiek: norminterpretaties, afwijkingsbesluiten en verificatie-evidence zijn de kern van hun governance. Die informatie hoort in een gecontroleerde omgeving, niet in een generiek platform waarvan je de data-flow niet kunt auditeren.

4. Leveranciersrisico: single point of failure

Een vierde risico dat zelden in de soevereiniteitsdiscussie wordt meegenomen is het **operationele leveranciersrisico** van volledige afhankelijkheid van één cloud-AI-provider. Wanneer je werkprocessen zijn gebouwd op de API van één partij, is die partij een single point of failure.

Dit is geen theoretisch risico. OpenAI faseerde in januari 2025 in één keer 33 modellen tegelijk uit. Bij grote modelupdates verandert de output op manieren die productiesystemen breken. Prijsverhogingen, gewijzigde gebruiksvoorwaarden of regionale beschikbaarheidsbeperkingen kunnen van de ene op de andere dag de basis onder een werkproces weghalen.

In de kritieke infrastructuur—waar continuïteit een kernvereiste is—is dit type afhankelijkheid onacceptabel. Soevereine architectuur met een verwisselbare provider layer elimineert dit risico: als één provider wegvalt of verandert, stap je over zonder herbouw.

Wat je niet ziet achter een cloud-API

Bij een cloud-API heeft de gebruiker geen zicht op wat er achter de schermen gebeurt. Providers voeren **silent updates** door—het model dat je vandaag aanroept kan morgen een andere versie, configuratie of optimalisatie draaien zonder dat je dat merkt. OpenAI faseerde in januari 2025 in één keer 33 modellen uit. Bij de lancering van GPT-5 brak de gewijzigde model-routing productiesystemen van klanten overnight.

Daarnaast passen providers routinematig **quantization** toe—een compressietechniek die de precisie van modelgewichten verlaagt om geheugen te besparen en meer gebruikers tegelijk te bedienen. Onderzoek (o.a. Red Hat, op basis van meer dan 500.000 evaluaties) toont aan dat 4-bit quantization tot 2–5% kwaliteitsverlies kan geven, afhankelijk van de taak. Batch-sizes, caching-strategieën en routing-keuzes beïnvloeden de output op manieren die de gebruiker niet kan zien en niet kan controleren.

Het resultaat: “dezelfde prompt” kan op verschillende momenten verschillende resultaten geven—niet door een fout van de gebruiker, maar door onzichtbare infrastructuurkeuzes van de provider. Voor generieke taken is dat acceptabel. Voor werkprocessen waar reproduceerbaarheid en traceability vereist zijn, is het een fundamenteel probleem.

Performance-degradatie bij piekbelasting

Een tweede operationeel risico van cloud-API’s is onvoorspelbare beschikbaarheid. Cloud-modellen draaien op gedeelde infrastructuur die miljoenen gelijktijdige gebruikers bedient. Tijdens piekuren—doorgaans de reguliere kantoortijden—kunnen responsetijden met 200–300% toenemen.

Wat soevereine AI concreet betekent voor de kritieke infrastructuur

Soevereine AI als term heeft betrekking op een set van architectuurkeuzes die bepalen hoeveel regie een organisatie heeft over haar AI-toepassing. In de context van kritieke infrastructuur vertaalt dat zich naar vijf dimensies:

1. Datasoevereiniteit

Projectdata wordt verwerkt en opgeslagen op infrastructuur binnen Nederlandse of Europese jurisdictie. Geen data-export naar cloudproviders voor verwerking. Geen retentie door derden.

2. Modelsoevereiniteit

De modellen die worden ingezet zijn controleerbaar (welke versie, welke weights, welke configuratie), vervangbaar en niet afhankelijk van één leverancier. Open-source modellen voor kritieke workloads maken het mogelijk om modelgedrag te auditeren en te reproduceren—essentieel wanneer AI-output in formele besluitvorming terechtkomt.

3. Infrastructuursoevereiniteit

De hardware waarop AI draait is eigendom van of volledig beheerd door de organisatie zelf, of door een partij onder Nederlandse of Europese jurisdictie. Geen afhankelijkheid van een buitenlandse partij die op afstand de dienst kan beëindigen of de functionaliteit kan wijzigen.

4. Portabiliteit

De architectuur is zo ontworpen dat een overstap naar een andere leverancier, een ander model of een andere hostingpartij mogelijk is zonder verlies van data, configuratie of functionaliteit. Geen vendor lock-in—niet in de LLM-laag, niet in de infrastructuurlaag, niet in de datastructuur.

5. Governance en traceerbaarheid

Elke AI-actie—van parsing tot classificatie tot antwoordgeneratie—wordt gelogd en is herleidbaar naar bron, model en methode. Dat maakt AI-output auditeerbaar, wat in regulated omgevingen niet optioneel is maar verplicht.

De opkomst van open-source modellen

De keuze voor open-source modellen is in 2026 geen concessie meer. Volgens Epoch AI lopen open-weight modellen nog maar circa drie tot zes maanden achter op de beste gesloten modellen—een kloof die eind 2023 nog 17,5 procentpunt bedroeg op MMLU. Op knowledge-benchmarks is de kloof nagenoeg nul; op reasoning en coding is het verschil single digits. Modellen als DeepSeek V3.2 en GLM-5, Qwen3.5, en Minimax presteren komen voor de meeste taken in de buurt van GPT-5 kwaliteit. De Stanford AI Index 2025 bevestigt deze convergentie.

Voor de kritieke infrastructuur betekent dit: je kunt kiezen voor volledige controle over je modellen zonder direct veel in te leveren op kwaliteit. De combinatie van open-source modellen, zelf gehost, met de mogelijkheid om te fine-tunen op domeindata, biedt een pad naar AI die zowel soeverein als hoogwaardig is.

Dimensie	Generieke cloud AI	Soevereine AI
Datalocatie	US/EU datacenter van provider, onder hun voorwaarden	NL/EU onder eigen jurisdictie, eigen hardware of trusted partner
Modelkeuze	Gesloten model, niet controleerbaar, silent updates	Open-source of swappable, auditeerbaar en reproduceerbaar
Infrastructuur	Gedeelde cloud, remote beheer door vendor	Eigen cluster of on-prem bij klant, volledig in eigen beheer
Portabiliteit	Hoge switchkosten, vendor lock-in	Provider-agnostisch ontwerp, exit altijd mogelijk
Audit trail	Beperkt of niet beschikbaar	Volledige provenance per AI-actie
Beschikbaarheid	Variabel, piekbelasting = degradatie	Voorspelbaar, eigen capaciteit

Secure by design: zes architectuurprincipes voor soevereine AI

Soevereiniteit is geen eigenschap die je achteraf toevoegt. Het is een ontwerpkeuze die in de architectuur zit. Bij predictIT hanteren we zes principes die samen de basis vormen voor soevereine AI in de kritieke infrastructuur.

1

Data blijft binnen de perimeter

Projectdata wordt verwerkt op eigen infrastructuur in Nederland. Geen API-calls met gevoelige data naar externe diensten. Waar compute wordt aangevuld, gebeurt dat op EU/NL-gehoste infrastructuur waarover wij of de klant zeggenschap hebben.

2

Eigen hardware, eigen regie

Onze kerninfrastructuur draait op hardware waarvan wij eigenaar zijn, gehost in Nederland. Daarnaast bieden we de mogelijkheid om on-prem te deployen in het datacenter van de klant. Geen afhankelijkheid van een derde partij die op afstand de dienst kan wijzigen of beëindigen.

3

Open-source modellen voor kritieke workloads

Voor parsing, indexering en beperkte reasoning op gevoelige data gebruiken we open-source modellen die we zelf hosten, inspecteren en aanpassen. Geen black box. Geen oncontroleerbare modelupdates die stilletjes de output veranderen.

4

Swappable provider layer

De architectuur is ontworpen met een verwisselbare LLM-laag. Klanten kunnen kiezen welk model ze inzetten, van welke provider, en op welke infrastructuur. Als een provider verandert, wegvalt of niet meer voldoet aan eisen: overstappen zonder herbouw.

5

Portability by design

Alle componenten—data, configuratie, modellen, workflows—zijn overdraagbaar. Geen proprietary formaten die lock-in creëren. De klant is eigenaar van de data en kan altijd weg. predictIT voorkomt te allen tijde een data lock-in als ontwerpprincipe.

6

Audit trail op elke AI-actie

Elke extractie, elke classificatie, elke gegenereerde analyse is herleidbaar naar bron, model, versie en methode. Dat maakt onze output bruikbaar in formele besluitvorming, verificatie en audit.

De paradox: partijen willen AI zelf doen, maar missen het datafundament

In de kritieke infrastructuur zien we een terugkerend patroon. IT-afdelingen en innovatieteams willen AI use cases zelf oppakken. Ze willen eigen copilots bouwen, eigen dashboards maken, eigen analyses draaien. Dat is begrijpelijk en dat moeten we als sector aanmoedigen.

Maar er is een fundamenteel obstakel: **een goede AI use case bestaat bij gratie van goede machine-readable data**. En die data is er bijna nooit. Projectdocumenten, PDF's, spreadsheets en contractbijlagen zijn rijk aan informatie, maar arm aan structuur. Zonder gestructureerde, gevalideerde data is elke AI-toepassing een huis op zand.

Dit is waar de rol van een gespecialiseerde parsing pipeline essentieel wordt: de vertaallaag van ongestructureerde projectdocumenten naar gevalideerde, semantisch verankerde asset data. Dit versnelt de AI use cases van interne teams die met deze stepping stone de ambitie mogelijk maakt.

De waarde van zo'n pipeline zit niet in de techniek van het parsen zelf—dat kan iedereen met een LLM-API proberen. De waarde zit in de "***domein-taste***": domein-jargon herkennen, synoniemen normaliseren, en een OTL (domeinontologie) toepassen bij het structureren van de output. Het verschil tussen "we hebben tekst geëxtraheerd" en "we hebben gevalideerde, semantisch verankerde asset data geproduceerd".

> **De kracht van specialistische modellen**
>
> Een fijngetuned open-source model van 7 miljard parameters, getraind op domeinspecifieke data, presteert voor specifieke taken vaak beter dan een generalistisch frontier-model—tegen een fractie van de kosten en draaiend op één GPU. De waarde zit niet in het grootste model, maar in het model dat jouw domein het beste begrijpt. Domein-taste plus fine-tuning op eigen data is de echte differentiator.

Onze visie hierop: als AI-output commodity wordt, maakt kwaliteit het verschil. En kwaliteit ontstaat door goed judgment. Wat weer ontstaat door diepgaande kennis van het toepassingsdomein. Het bij elkaar brengen van domeinexpertise en AI-expertise is hierbij de kern.

> **De soevereine dimensie:** soevereiniteit en domeinexpertise zijn hier onlosmakelijk verbonden: je kunt de data niet goed structureren als je het domein niet begrijpt, en je kunt de data niet verantwoord verwerken als je de soevereiniteit niet borgt.

Hoe predictIT soevereine AI in de praktijk brengt

1. Infrastructuur: eigen cluster, eigen hardware

Onze kerninfrastructuur draait op een eigen cluster in Nederland, op hardware waarvan wij eigenaar zijn. Voor kritieke AI-workloads beschikt predictIT over eigen GPU's. Voor extensieve GPU-intensieve workloads vullen we aan met EU/NL-gehoste compute. De klant weet altijd waar de data is en wie er toegang toe heeft.

2. Portability voor migratie naar eigen infrastructuur

Indien gewenst wordt de mogelijkheid geboden om de volledige stack te deployen in het eigen datacenter van de klant—als het nodig is zelfs volledig air-gapped.

3. Governance, monitoring en logging

AI-services krijgen dezelfde governance als menselijke gebruikers, daarmee kan met role-based access control exact toegekend worden welke service welke rechten krijgt. Alle lees- en schrijfoperaties worden gelogd voor traceerbaarheid en auditeerbaarheid.

4. Modellen: open-source kern met swappable provider layer

Voor kritieke workloads—parsing, indexering, classificatie van gevoelige projectdata—gebruiken we open-source modellen die we zelf hosten en beheren. Dat betekent: gecontroleerde modelupdates, grip op de output en de exacte configuratie. We kunnen meebewegen met de toekomstige ontwikkelingen en slim kiezen wanneer grote en wanneer kleine modellen in te zetten.

5. Verwisselbare provider layer

Voor niet-gevoelige of generieke taken kan de klant kiezen welk commercieel model via welke provider wordt ingezet.

6. Grip op de hele keten als basis voor innovatie

Wanneer je de hele AI-keten beheerst—van gebruikersinterface tot modelkeuze tot hardware—ontstaat een unieke positie voor continue verbetering. Je kunt:

- modellen wisselen zonder migratie of downtime
- fine-tunen op domeindata voor betere resultaten op specifieke taken
- benchmarken op eigen data voordat je naar productie gaat
- experimenteren met kleinere, snellere modellen voor specifieke workloads
- nieuwe open-source modellen evalueren zodra ze beschikbaar komen—zonder afhankelijkheid van de roadmap van een enkele provider

> **Die innovatievrijheid heb je niet als je vast zit aan een cloud-API waarvan je de binnenkant niet kent. Grip op de hele keten is niet alleen een veiligheidskeuze—het is een strategische keuze voor toekomstige wendbaarheid.**

Begrippenkader

Soevereine AI	AI-systemen waarbij een organisatie volledige regie heeft over data, modellen, infrastructuur en governance—inclusief de mogelijkheid om van leverancier te wisselen.
NIS2	EU-richtlijn (2022/2555) voor netwerk- en informatiebeveiliging. Classificeert energie, water, transport en digitale infrastructuur als “essentiële entiteiten” met verplichtingen rondom risicobeheer en incidentmelding.
EU AI Act	Europese verordening (2024/1689) die AI-systemen classificeert op risico. High-risk systemen moeten per 2 augustus 2026 voldoen aan eisen voor documentatie, conformiteit en menselijk toezicht.
CLOUD Act	US-wetgeving (2018) die Amerikaanse autoriteiten de bevoegdheid geeft data op te vragen bij US-gevestigde cloudproviders, ongeacht waar die data fysiek is opgeslagen.
OTL	Object Type Library—een gestandaardiseerde domeinontologie met objecttypen, relaties en definities. Fungeert als semantische ruggengraat voor lifecycle-traceability en data-governance.
Quantization	Compressietechniek die de precisie van modelgewichten verlaagt (bijv. van 16-bit naar 4-bit) om geheugen te besparen. Kan meetbaar kwaliteitsverlies veroorzaken, afhankelijk van de taak.
Portability by design	Architectuurprincipe waarbij alle componenten (data, modellen, configuratie) overdraagbaar zijn naar een andere leverancier of omgeving, zonder verlies van functionaliteit.
Swappable provider layer	Architectuurpatroon waarbij het LLM als verwisselbaar component wordt behandeld: de klant kiest welk model, van welke provider, op welke infrastructuur.
Fine-tuning	Het doortrainen van een bestaand model op domeinspecifieke data, waardoor een kleiner model voor specifieke taken beter kan presteren dan een groter generalistisch model.