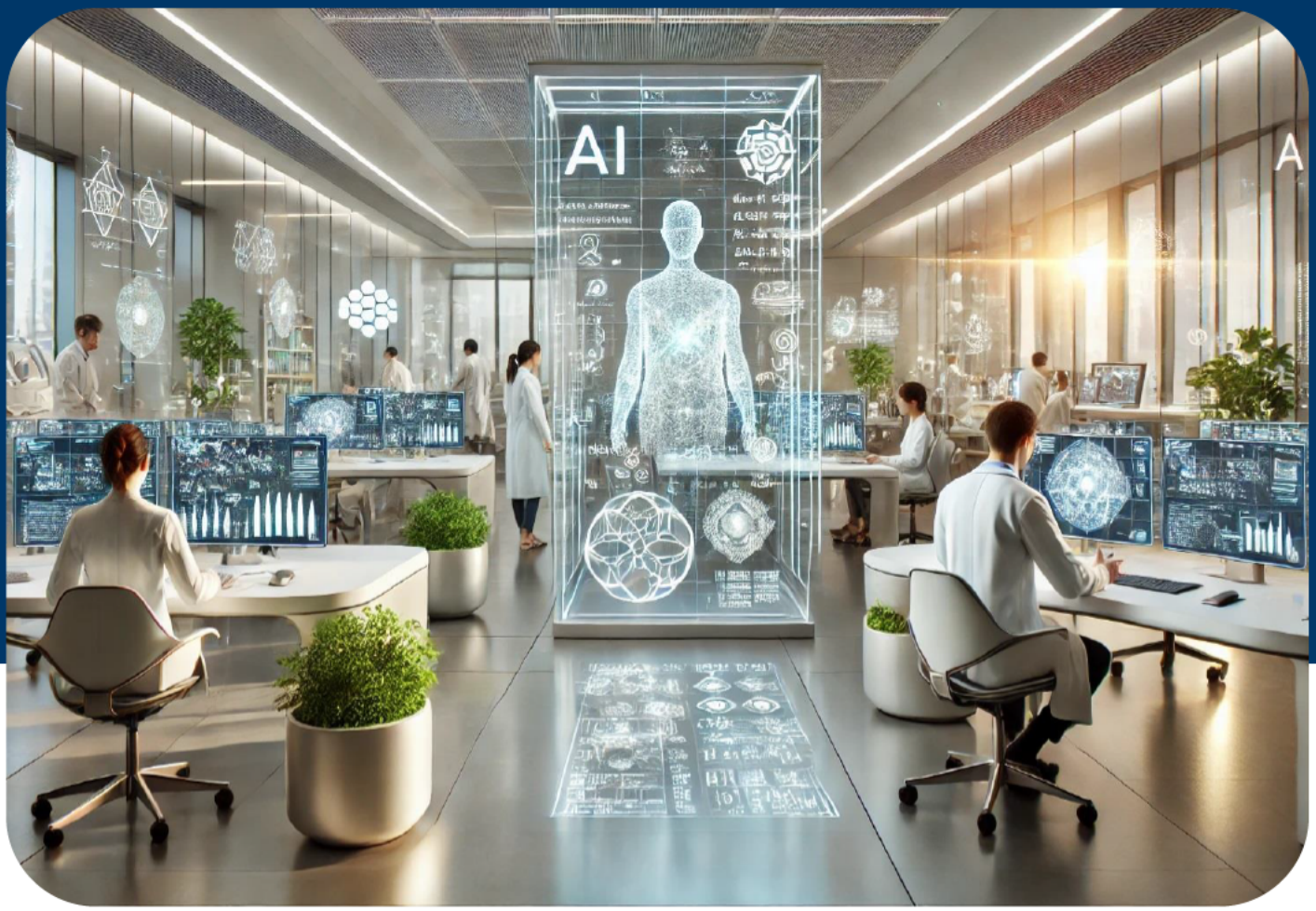
dialogic

Onderzoek voor *onderbouwd* beleid.



Meta-analyse

Consequenties van beleidsopties voor een AI-fabriek

Auteurs

Dr. Max Kemman

Ir. Arthur Vankan

Ir. Menno Driesse

Guido de Moor MSc MA

Ir. Tommy van der Vorst

Nino van Sambeek MSc

Meta-analyse

Consequenties van beleidsopties voor een AI-fabriek

Auteurs

Dr. Max Kemman
Ir. Arthur Vankan
Ir. Menno Driesse
Guido de Moor MSc MA
Ir. Tommy van der Vorst
Nino van Sambeek MSc

Opdrachtgever

Ministerie van Onderwijs, Cultuur en Wetenschap (OCW)

Publicatienummer

2024.131-2436

Datum

29 november 2024

Beeld omslag

DALL-E

Inhoud

Samenvatting

Kunstmatige intelligentie (AI) is een belangrijke technologie voor het (toekomstige) verdienvermogen en de welvaart van Nederland, met toepassingen in de wetenschap, het bedrijfsleven en bij de overheid. Het kabinetsbeleid is gericht op het ontwikkelen van een sterk AI-ecosysteem in Nederland. Op Europees niveau wordt beleidsmatig ingezet op het realiseren van rekencapaciteit om een mondiale koploperpositie te verkrijgen op het gebied van AI. Vanuit EuroHPC Joint Undertaking loopt inmiddels een call om AI-fabrieken te realiseren op basis van cofinanciering. Met een AI-fabriek wordt bedoeld:

> *Een gecentraliseerde of gedistribueerde entiteit die een diensteninfrastructuur voor supercomputing op het gebied van AI aanbiedt die bestaat uit een voor AI geoptimaliseerde supercomputer of een AI-gericht deel van een supercomputer, een bijbehorend datacentrum, specifieke toegang en AI-georiënteerde supercomputingdiensten, en die talent aantrekt en bundelt om de competenties te leveren die nodig zijn om supercomputers voor AI te gebruiken.*

Naar aanleiding van de ontwikkelingen op het gebied van AI en met het oog op de call vanuit EuroHPC JU hebben de ministeries van OCW, EZ en BZK een gezamenlijke verkenning aangekondigd op basis van drie beleidsopties:

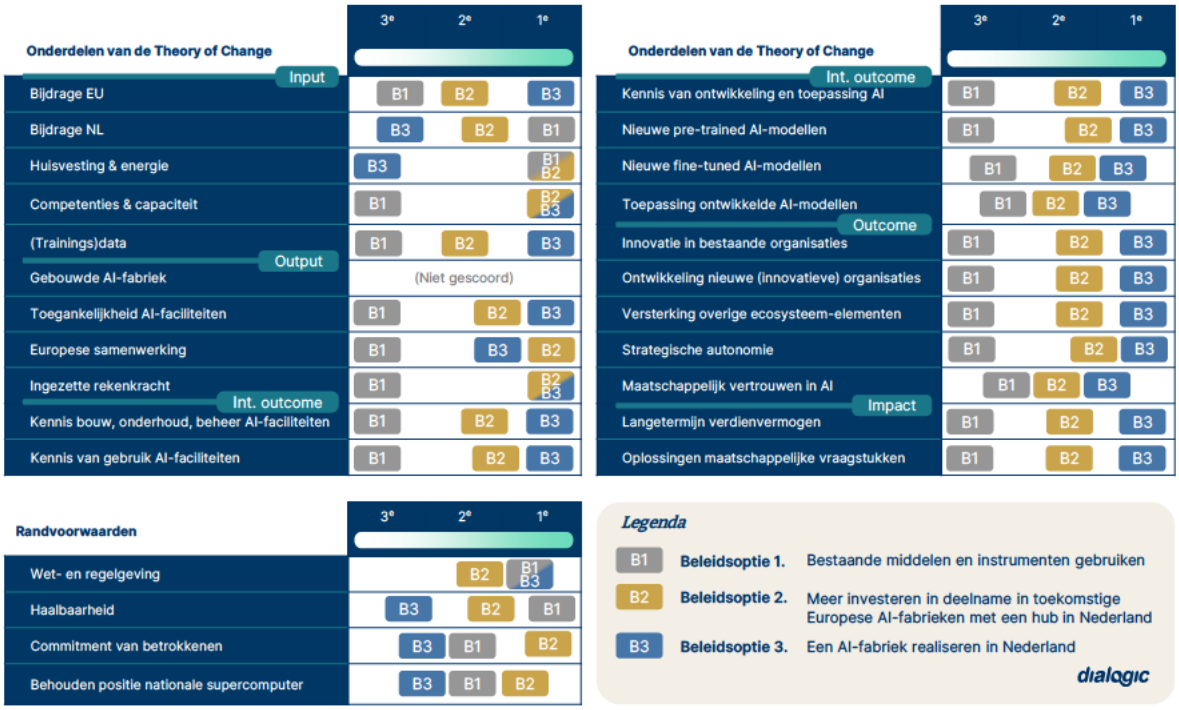
- B1. Bestaande middelen en instrumenten gebruiken.
- B2. Meer investeren in deelname in toekomstige Europese AI-fabrieken met een hub in Nederland.
- B3. Een AI-fabriek realiseren in Nederland.

Om nut en noodzaak van een AI-fabriek te verkennen, zijn door verschillende actoren in het systeem drie cases uitgewerkt: een **wetenschapscasus** gericht op de wetenschappelijke meerwaarde (uitgewerkt door AI-onderzoekers), een **publieke sector casus** gericht op de meerwaarde voor de Nederlandse overheid en dienstverlening (uitgewerkt door BZK), en een **bedrijfscasus** gericht op de economische meerwaarde (uitgewerkt door AiNed).

Dialogic is gevraagd om een **meta-analyse** om de te maken keuze op basis van de drie beleidsopties te onderbouwen. Voor de meta-analyse hebben we een **analysekader** ontwikkeld waarin criteria, randvoorwaarden en flankerend beleid ten aanzien van het AI-ecosysteem en een AI-fabriek in Nederland zijn uitgewerkt. De meta-analyse is vervolgens uitgevoerd door de drie **cases** te beoordelen met het analysekader.
Gedurende het proces zijn daarnaast gesprekken gevoerd met de auteurs van de cases en zijn **witte vlekken** ingevuld met enkele gesprekken met stakeholders uit het AI-ecosysteem.

Tabel S1 toont een overzicht van de resultaten van de meta-analyse. Per criterium maken we een ranking tussen de beleidsopties. We doen geen uitspraken over onderlinge afwegingen tussen criteria.

Tabel S1. Overzicht resultaten meta-analyse: ranking van beleidsopties per onderdeel van de theory of change en randvoorwaarden



We geven de volgende aandachtspunten mee voor het vervolg:

- Beleidsoptie B2 heeft de grootste waarde bij een serieuze investering in de hub en in het internationale consortium. B2 kan in theorie geïmplementeerd worden met een minimale investering, maar veel van de voordelen en opbrengsten die we in kaart brengen in dit rapport vervallen dan. Gesteld kan worden dat bij een minimale investering in B2, deze nagenoeg samenvalt met B1. Met een grote investering in B2 valt deze meer samen met B3 (meer expertise en meer rekenkracht, zonder de risico's van B3).
- De toegankelijkheid (de hoeveelheid rekentijd en de snelheid waarmee een aanvraag kan worden toegekend en ingepland) van een AI-fabriek met beleidsoptie B2 is sterk afhankelijk van de afspraken die worden gemaakt in het internationaal consortium. De mate waarin B2 minder goed scoort op toegankelijkheid dan B3 is dus enigszins onzeker en afhankelijk van de afspraken die gemaakt moeten worden binnen B2.
- Beleidsoptie B3 biedt de grootste mogelijke opbrengsten, maar brengt ook de grootste risico's met zich mee. Het is de vraag of de meerwaarde van B3 ten opzichte van B2 en de benodigde middelen voor B3 daarmee doelmatig is en blijft. Daarentegen is het niet altijd mogelijk om R&D op volledig doelmatige wijze te stimuleren.
- Als er middelen worden gereserveerd voor de implementatie van beleidsopties B2 of B3 worden idealiter niet alleen incidentele middelen vrij gemaakt om te voldoen aan de EuroHPC JU-call, maar wordt een meerjarenbegroting opgesteld voor afschrijving, beheer en vervanging van AI-faciliteiten. Binnen een dergelijke

meerjarenbegroting kan ook aandacht worden besteed aan de (opvolging van de) nationale supercomputer en andere digitale infrastructuur.

- Om methodologische redenen zijn de drie beleidsopties afzonderlijk geanalyseerd. Een combinatie van (onderdelen van) beleidsopties biedt mogelijk kansen om sterktes te combineren en risico's of zwaktes te mitigeren.
- De meerwaarde van een AI-fabriek voor het AI-ecosysteem wordt ten dele bepaald door imagovorming van Nederland als hightech land. Beleidsopties B2 (met een hub) en B3 (met een eigen faciliteit) bieden kansen.

Tabel S2, op de volgende pagina, toont een overzicht van de concrete verschillen tussen de beleidsopties op rekenkracht, data en expertise. Hierbij merken we op dat de uiteindelijke effecten afhankelijk zijn van de implementatie.

Tabel S2. Overzicht resultaten meta-analyse: ranking van beleidsopties per onderdeel van de theory of change en randvoorwaarden

	B1		B2		B3	
Kenmerk	**Score**	**Uitleg**	**Score**	**Uitleg**	**Score**	**Uitleg**
Reken-kracht	**Laag**	Er komt geen extra rekenkracht beschikbaar binnen Nederland. Europees kan in concurrentie rekenkracht worden aangevraagd bij andere AI-fabrieken.	**Medium / Hoog**	Er komt extra rekenkracht beschikbaar voor Nederland. Er komt mogelijk meer rekenkracht beschikbaar dan met B3 door schaalvoordelen van samenwerking in een internationaal consortium.	**Hoog**	Er komt extra rekenkracht beschikbaar voor Nederland. Een Nederlandse partij beheert de rekenkracht en prioriteert aanvragen. In het geval van een 'kleine' AI-fabriek kan de score op **Medium** uitkomen, door minder beschikbare rekenkracht.
Data	**Laag**	Er komt geen datacentrum beschikbaar waar grootschalige datasets kunnen worden opgeslagen, en verwerkt voor/door AI-modellen. Data dient bij commerciële aanbieders te worden verzameld en verwerkt, of bij AI-fabrieken elders in Europa.	**Medium**	Er komt een datacentrum beschikbaar waar grootschalige datasets kunnen worden verwerkt voor/door AI-modellen. Daardoor is geen afhankelijkheid van commerciële aanbieders voor AI-verwerking van data. Bepaalde use cases worden niet ondersteund doordat data niet buiten de landgrenzen mag worden verwerkt.	**Hoog**	Er komt een datacentrum beschikbaar waar grootschalige datasets kunnen worden verwerkt voor/door AI-modellen. Daardoor is er geen strategische afhankelijkheid van commerciële aanbieders of buitenlandse AI-fabrieken voor AI-verwerking van data. Waarschijnlijk worden niet alle use cases ondersteund doordat data niet buiten de muren van instellingen mag worden verwerkt, maar een AI-fabriek binnen Nederland maakt het mogelijk hierover het gesprek aan te gaan.
Expertise	**Laag**	Er wordt geen extra expertise opgebouwd ten aanzien van AI-faciliteiten of ontwikkeling en toepassing van AI-modellen.	**Hoog**	Er wordt een hub opgebouwd met extra expertise ten aanzien van ontwikkeling en toepassing van AI-modellen. Dit is de hoogwaardige expertise die het meest van belang is voor het AI-ecosysteem. In het geval dat alleen wordt aangesloten op een internationaal consortium zonder investeringen in een nationale hub kan de score op **Laag** uitkomen.	**Hoog**	Er wordt een hub opgebouwd met extra expertise ten aanzien van ontwikkeling en toepassing van AI-modellen. Ook wordt er expertise opgebouwd ten aanzien van het bouwen en beheren van AI-infrastructuur.

Begrippenlijst

Begrip	Toelichting
Algoritme	Een verzameling van regels en instructies die door een computer kunnen worden uitgevoerd. Algoritmes kunnen worden toegepast op informatie en deze combineren en analyseren, om zo tot nieuwe informatie te komen.
AI	*Kunstmatige intelligentie.* Het door een machine laten uitvoeren van intelligente taken die voordien waren voorbehouden aan mensen.
Generatieve AI	Een vorm van AI waarmee op instructie van de gebruiker tekst, beeld en andere inhoud kunnen worden gegenereerd, en die zeer breed inzetbaar zijn. Een bekend voorbeeld van generatieve AI is ChatGPT.
AI-model	Een wiskundig/statistisch model dat kan worden gebruikt om AI te realiseren. Een AI-model bestaat uit een groot aantal 'parameters' (getallen) waarin de 'kennis' van de AI besloten ligt, en een (wiskundig) algoritme om met deze parameters en invoer van de gebruiker tot een bepaalde uitvoer te komen (bijvoorbeeld een voorspelling of antwoord op een vraag).
Taalmodel	Een AI-model dat werkt op basis van taal – het model accepteert tekst als invoer (bijvoorbeeld een vraag of instructie) en produceert tekst als uitvoer (het antwoord op de vraag). Een (kleinschalig) voorbeeld van een taalmodel is het autocorrect-model dat voorspelt welk volgend woord je wil typen op een smartphone.
Groot taalmodel (LLM)	Een *groot* taalmodel (LLM: Large Language Model) is een taalmodel dat bestaat uit een zeer groot aantal parameters (enkele miljarden) en is gebaseerd op basis van zeer grote hoeveelheden tekst. Hierdoor lijkt het model over veel algemene kennis te beschikken. Een voorbeeld van een groot taalmodel is GPT-4.
Foundation model	Een AI-model dat geschikt is voor een breed scala aan taken. Foundation modellen zijn over het algemeen gebaseerd op zeer grote hoeveelheden gegevens en (lijken te) beschikken over brede algemene kennis. Een foundation model bestaat meestal uit zeer veel parameters en is zeer kostbaar om te ontwikkelen. Een voorbeeld van een foundation (taal)model is GPT-4.

1 Inleiding

In dit rapport worden drie beleidsopties voor het eventueel realiseren van een AI-fabriek ten behoeve van een sterk AI-ecosysteem in Nederland geanalyseerd.

Leeswijzer

In dit hoofdstuk beschrijven we de aanleiding voor de meta-analyse en de aanpak. In hoofdstuk 2 werken we uit wat een AI-fabriek is, waarvoor deze kan worden ingezet en welke plek een AI-fabriek kan spelen in het AI-ecosysteem. In hoofdstuk 3 vatten we de drie cases samen waarvoor een AI-fabriek van meerwaarde kan zijn; wetenschap, overheid en bedrijfsleven. In hoofdstuk 4 introduceren we het analysekader en bespreken we de uitkomsten van de meta-analyse. In hoofdstuk 5 presenteren we onze conclusies.

1.1 Aanleiding voor de meta-analyse

Kunstmatige intelligentie (AI) is een belangrijke technologie voor het (toekomstige) verdienvermogen en de welvaart van Nederland [1]. De introductie van *generatieve* AI bracht de ontwikkeling en toepassing van AI in een stroomversnelling. Generatieve AI is een vorm van kunstmatige intelligentie die in staat is om tekst en beeld te genereren, en zeer breed inzetbaar is. Naast generatieve AI bestaan ook andere zeer waardevolle en kansrijke vormen van AI. Beide kennen toepassingen in de wetenschap, het bedrijfsleven en bij de overheid.

In januari 2024 bracht het kabinet de *Overheidsbrede visie op generatieve AI* uit [2]. In deze visie wordt de verwachte enorme impact van generatieve AI op verschillende sectoren en domeinen binnen de samenleving benadrukt, zowel in positieve als negatieve zin. De overheidsvisie erkent de potentie van generatieve AI om maatschappelijke en wetenschappelijke vraagstukken op te lossen en productiviteit te verhogen, maar wijst ook op de risico's zoals desinformatie, privacy-problemen, en ongelijkheid. Met de visie wil het kabinet een kader scheppen waarin generatieve AI op een verantwoorde manier wordt ontwikkeld en toegepast, met oog voor publieke waarden zoals veiligheid, rechtvaardigheid, en duurzaamheid. Dit omvat onder andere het stimuleren van innovatie, het versterken van regelgeving, en het vergroten van kennis en kunde rondom AI. De Rijksbrede visie op generatieve AI benadrukt de noodzaak van samenwerking op nationaal en internationaal niveau om een sterk AI-ecosysteem te creëren dat in lijn is met Europese waarden en normen.

Het kabinetsbeleid, dat is gebaseerd op de Rijksbrede visie op generatieve AI en diverse andere aanpalende beleidsstukken,[1] is gericht op het ontwikkelen van een **sterk AI-ecosysteem in Nederland**. Toegang tot voldoende en geschikte rekenkracht is hier een essentieel onderdeel van, zowel voor generatieve als niet-generatieve AI. Hoewel een deel van de benodigde rekenkracht door de markt wordt geleverd, wordt ook onderkend dat ook de overheid een rol heeft in het voorzien hierin. Ook op Europees niveau wordt beleidsmatig ingezet op het realiseren van rekencapaciteit om een mondiale koploperpositie te verkrijgen op het gebied van AI [3]. Vanuit EuroHPC JU (*Joint Undertaking*) loopt inmiddels een competitieve call om AI-fabrieken te realiseren op basis van cofinanciering.

Naar aanleiding van de ontwikkelingen op het gebied van AI en met het oog op de call vanuit EuroHPC JU hebben de ministeries van OCW, EZ en BZK een gezamenlijke verkenning aangekondigd naar nut en noodzaak van een AI-faciliteit [1]. In de verkenning worden, met het oog op de EuroHPC JU-call, drie beleidsopties uitgewerkt die gaan over de vraag hoe Nederland het beste de toegang tot rekenkracht en daarmee het versterken van het AI-ecosysteem kan vormgeven. De drie beleidsopties zijn de volgende:

B1. Bestaande middelen en instrumenten gebruiken – onderzoekers met AI-vraagstukken die grote hoeveelheden rekenkracht nodig hebben, maken gebruik de bestaande nationale supercomputer *Snellius*, van capaciteit die binnen het EuroHPC-consortium beschikbaar is in andere landen, of van commercieel beschikbaar rekencapaciteit van onder andere cloudaanbieders.

Benodigde investering: geen additionele middelen.

B2. Meer investeren in deelname in toekomstige Europese AI-fabrieken met een hub in Nederland. Nederland is al met twee miljoen euro inleg partner in de Finse LUMI-supercomputer, en heeft een uitnodiging ontvangen om deel te nemen in de opschaling van LUMI. Daarnaast investeert Nederland al acht miljoen euro in het Franse Jules Verne-consortium, dat werkt aan de 'Alice Recogue'-supercomputer. Met 'hub' wordt verwezen naar een lokaal onderdeel zoals een kenniscluster, al dan niet op een specifieke fysieke locatie, die de toegang en inzet van de faciliteiten elders faciliteert.[2]

[1] Het Nederlandse beleid op het terrein van AI wordt onder andere beschreven in het Strategisch Actieplan voor Artificiële Intelligentie (SAPAI), de Kabinetsreactie Wetenschappelijke Raad voor het Regeringsbeleid (WRR) rapport Opgave AI, de Strategie Digitale Economie (SDE), de Werkagenda Waardengedreven Digitaliseren, de Agenda Digitale Open Strategische Autonomie (DOSA), de Nationale Technologiestrategie (NTS) en de Overheidsbrede visie Generatieve AI.

[2] In de eerdere aankondiging van de verkenning werd de hub nog niet genoemd. Deze invulling van beleidsoptie B2 is daarmee een evolutie op basis van voortschrijdend inzicht bij de desbetreffende ministeries.

Benodigde investering: minimaal vijf tot maximaal tientallen miljoenen euro (hoe groter de bijdrage, hoe groter het Nederlandse aandeel in de rekencapaciteit/tijd).

B3. Een AI-fabriek realiseren in Nederland. Deze beleidsoptie behelst het realiseren van een volledige AI-fabriek – inclusief de onderliggende HPC-faciliteit, op Nederlandse bodem. Deze faciliteit wordt onderdeel van het bredere Europese supercomputerecosysteem (via EuroHPC). Omdat er sprake is van cofinanciering vanuit EuroHPC JU wordt de helft van de rekencapaciteit beschikbaar gemaakt voor gebruikers uit andere Europese landen.

Benodigde investering: Maximaal 215 miljoen euro aan initiële Nederlandse investeringen en daaropvolgende toekomstige kosten.

Om nut en noodzaak van een AI-fabriek te verkennen, zijn door verschillende actoren in het systeem drie cases uitgewerkt: een **wetenschapscasus** gericht op de wetenschappelijke meerwaarde (uitgewerkt door AI-onderzoekers), een **publieke sector casus** gericht op de meerwaarde voor de Nederlandse overheid en dienstverlening (uitgewerkt door BZK), en een **bedrijfscasus** gericht op de economische meerwaarde (uitgewerkt door AiNed). Deze cases lichten we nader toe in hoofdstuk 3.

1.2 Aanpak

Dialogic is gevraagd om een **meta-analyse** om de te maken keuze op basis van de drie beleidsopties te onderbouwen. Deze meta-analyse moet de beleidsopties zoals uitgewerkt in de cases toetsen aan de hand van onderbouwde criteria, randvoorwaarden en flankerend beleid. Bij deze meta-analyse zijn een tweetal overwegingen van belang, die voortkomen uit de afbakening van de opdracht door de opdrachtgever en praktische overwegingen:

1. **De eerder benoemde beleidsopties zijn het startpunt van de meta-analyse.** De drie beleidsopties worden als gegeven gezien. We verkennen geen alternatieve beleidsopties of (concrete) subbeleidsopties. Wel kunnen aandachtspunten binnen de gegeven beleidsopties uit het onderzoek komen. Ook is het mogelijk dat beleidsopties in combinatie worden uitgevoerd.
2. **We presenteren per criterium van het analysekader welke beleidsoptie de meeste meerwaarde genereert.** De drie beleidsopties leiden tot verschillende financiële, wetenschappelijke, economische en publieke consequenties. Om een beleidskeuze te onderbouwen is een precieze bepaling van de orde van grootte (bijv. van economische impact) niet benodigd, maar volstaat een ranking tussen de beleidsopties. We doen daarbij geen uitspraken over onderlinge afwegingen tussen criteria (welke criteria de doorslag geven voor een uiteindelijke beleidskeuze), maar we wegen alle beleidsopties op alle criteria.

Voor de meta-analyse hebben we een **analysekader** ontwikkeld waarin criteria, randvoorwaarden en flankerend beleid ten aanzien van het AI-ecosysteem en een AI-fabriek in Nederland zijn uitgewerkt in een *theory of change*. Dit analysekader is besproken met de opdrachtgevers en auteurs van de cases. De meta-analyse is vervolgens uitgevoerd door de drie **cases** te beoordelen met het analysekader. Gedurende het proces zijn daarnaast gesprekken gevoerd met de auteurs van de cases en zijn **witte vlekken** ingevuld met enkele gesprekken met stakeholders uit het AI-ecosysteem. Concreet is gesproken met vertegenwoordigers van SURF, AiNed, NOLAI, NWO, eScience Center, Koninklijke Bibliotheek, VNO-NCW, en AI-onderzoekers.

2 Kenmerken van een AI-fabriek

In dit hoofdstuk werken we concreet uit wat het onderwerp is van de meta-analyse. In paragraaf 2.1 bespreken we de concrete onderdelen van een AI-fabriek, waaronder rekenkracht, data en expertise. In paragraaf 2.2 lichten we toe voor welke AI-taken een AI-fabriek ingezet kan worden. In paragraaf 2.3 werken we ten slotte uit wat een AI-ecosysteem is en welke plek een AI-fabriek hierin inneemt.

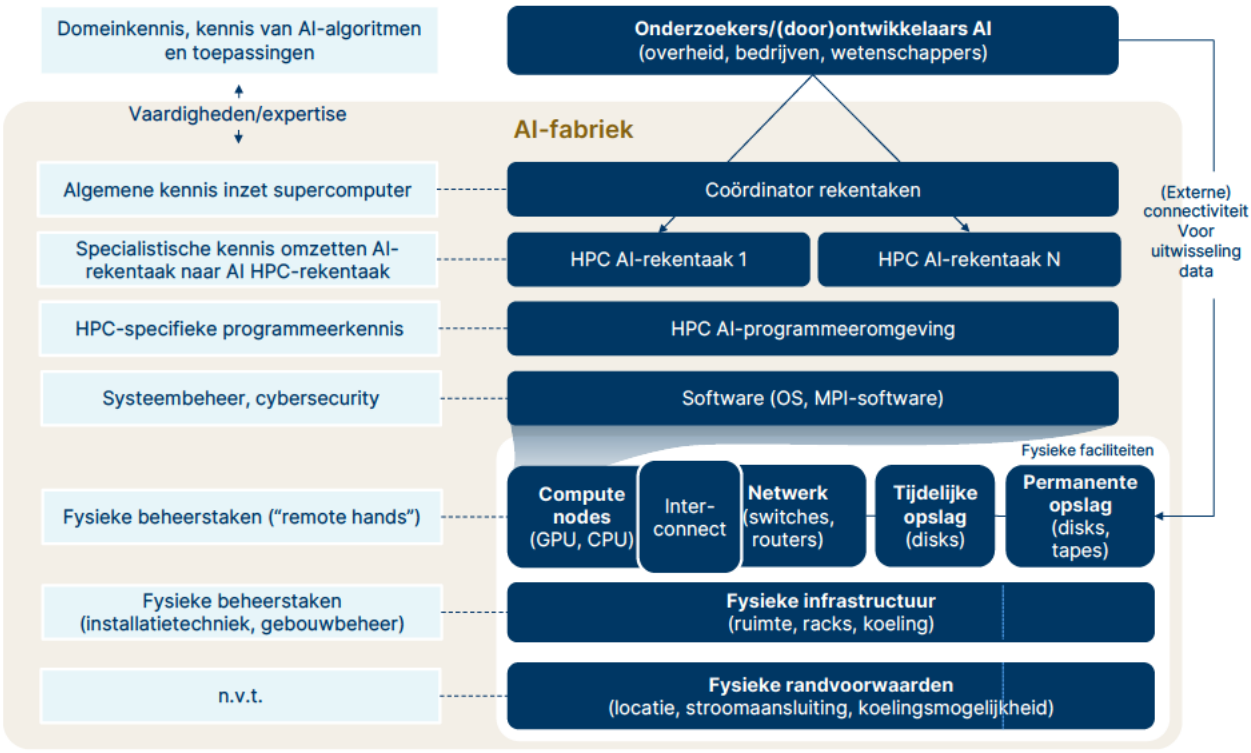
2.1 Onderdelen van een AI-fabriek

EuroHPC JU (*Joint Undertaking*) is een gezamenlijk initiatief van de Europese Commissie, Europese landen en private partijen om een supercomputer-ecosysteem van wereldklasse te ontwikkelen in Europa. Vanuit EuroHPC JU is een call opgestart waarmee financiële ondersteuning kan worden verkregen voor het realiseren van een AI-fabriek [3]. Met een AI-fabriek wordt in de gewijzigde verordening bedoeld:

> ***Een gecentraliseerde of gedistribueerde entiteit die een diensteninfrastructuur voor supercomputing op het gebied van AI aanbiedt die bestaat uit een voor AI geoptimaliseerde supercomputer of een AI-gericht deel van een supercomputer, een bijbehorend datacentrum, specifieke toegang en AI-georiënteerde supercomputingdiensten, en die talent aantrekt en bundelt om de competenties te leveren die nodig zijn om supercomputers voor AI te gebruiken.*** [4]

Met “voor AI geoptimaliseerde supercomputer” wordt een supercomputer bedoeld die in de eerste plaats ontworpen is voor het trainen van grootschalige AI-modellen voor algemene doeleinden, en het ontwikkelen van opkomende AI-toepassingen, maar die ook voor andere (niet-AI) toepassingen gebruikt kan worden.

We onderscheiden in de AI-fabriek vier fundamentele ingrediënten: (1) de rekenkracht van de supercomputer, (2) grootschalige opslagcapaciteit, (3) het datacentrum (fysieke faciliteiten), en (4) kennis en expertise. Deze ingrediënten hangen nauw samen en bestaan uit verschillende lagen die op elkaar bouwen, zoals weergeven in Figuur 1.



Figuur 1. Onderdelen van een AI-fabriek: expertise en kennis (links) en concrete faciliteiten (rechts). Bron: Dialogic o.b.v. literatuuranalyse en gesprekken

De drie beleidsopties kunnen als volgt in de figuur gelezen worden:

- **Beleidsoptie B1** ('bestaande middelen') behoudt de onderzoekers/(door)ontwikkelaars van AI, maar er komt geen (door Nederlandse partijen gefaciliteerde toegang tot een) AI-fabriek beschikbaar binnen Nederland.
- **Beleidsoptie B2** ('Europese deelname') voegt aan B1 de vier onderdelen (boven het blok 'Fysieke faciliteiten') van een AI-fabriek toe. Er wordt geïnvesteerd in een hub waarin expertise wordt opgebouwd rondom het ontwikkelen van AI. Afhankelijk van de invulling van de hub kan de expertise variëren van domeinkennis op specifieke AI-toepassingen bij onderzoekers/(door)ontwikkelaars, tot meer operationele expertise zoals het opzetten van AI HPC-rekentaken, programmeren op HPC en systeembeheer om een AI-faciliteit te gebruiken bij coördinatoren. De coördinatie van rekentaken geschiedt (voor de 'Nederlandse' rekencapaciteit) in Nederland, en voor de totale rekencapaciteit die beschikbaar is in het consortium. De rekencapaciteit (de supercomputer) staat fysiek in een ander land, bijvoorbeeld LUMI in Finland.
- **Beleidsoptie B3** ('Nederlandse AI-fabriek') voegt aan B2 de fysieke faciliteiten van een AI-fabriek toe op Nederlandse bodem.

Hieronder werken we de figuur in meer detail uit en lichten we de verschillende onderdelen van de AI-fabriek toe.

Aan het begin van de keten zien we de **onderzoekers en (door)ontwikkelaars** van AI-algoritmes. Dit zijn onder meer data scientists die werkzaam zijn bij wetenschappelijke instellingen, start-ups, mkb, grotere bedrijven of de overheid. Deze onderzoekers en ontwikkelaars hebben kennis van AI-algoritmes, een concrete wens om deze toe te passen, en daaruit een noodzaak voor rekencapaciteit en topkennis over modellen en tools.

2.1.1 Concrete faciliteiten

Aan de rechterkant van de figuur vinden we de tastbare faciliteiten die samen een AI-fabriek vormen. Kenmerkend hierbij is de gelaagdheid. Bij een AI-fabriek ontstaat uiteraard allereerst, van onder naar boven kijkend, het beeld van een **fysieke faciliteit** – een gebouw waarin de apparatuur en mensen (kennis) aanwezig is. Deze fysieke faciliteit heeft bepaalde **fysieke randvoorwaarden**, zoals een stroomaansluiting en koelingsfaciliteit (denk bijvoorbeeld aan een nabijgelegen water). De volgende laag wordt gevormd door de **hardware** die nodig is – hierbij gaat het allereerst om de *compute nodes*. Dit zijn computers die specifiek zijn ontworpen voor gebruik in een datacenter (*servers*), en de 'ruwe rekenkracht' bevatten (in de vorm van CPU's en GPU's).

CPU's en GPU's

Een CPU (Central Processing Unit) is een chip die grote hoeveelheden berekeningen in serie kan uitvoeren. CPU's zijn te vinden in praktisch ieder digitaal apparaat en variëren in de snelheid waarmee berekeningen kunnen worden uitgevoerd. Een GPU (Graphics Processing Unit) is een chip die is gespecialiseerd in het uitvoeren van grote hoeveelheden (parallelle) berekeningen die nodig zijn bij (onder andere) grafische toepassingen, zoals het genereren van 3D-beelden in videogames. Voor het trainen en gebruiken van AI-algoritmes zijn dezelfde soort berekeningen nodig. Hoewel een CPU deze berekeningen ook zou kunnen uitvoeren, kan een GPU dit veel sneller, doordat het de rekentaak op de chip kan opdelen in een groot aantal relatief eenvoudige berekeningen, en deze parallel kan uitvoeren.[3]

Het Amerikaanse NVIDIA is de marktleider op het gebied van GPU's. [5] NVIDIA ontwikkelt en levert de chips (fabricage hiervan wordt door TSMC gedaan) en de bijbehorende softwarebibliotheken (CUDA). Deze bevat onder andere geoptimaliseerde routines voor het uitvoeren van berekeningen die veelgebruikt zijn bij AI. Zowel NVIDIA zelf als derde partijen (OEM's) leveren insteekkaarten of modules (de datacenterversie van een 'videokaart') of geïntegreerde systemen op basis van de chips en ontwerpen van NVIDIA.

De belangrijkste concurrent van NVIDIA in deze markt is AMD – deze levert vergelijkbare hardware (ook qua ruwe rekenpresentaties), maar het software-ecosysteem

[3] Voor een illustratie van het verschil, zie [youtube.com]

rondom AMD GPU's voor machine learning is minder sterk ontwikkeld. Veruit de meeste machine learning-raamwerken ondersteunen alleen CUDA of zijn alleen goed geoptimaliseerd voor CUDA. De huidige LUMI-faciliteit bijvoorbeeld gebruikt AMD-chips. In ten minste één van de geanalyseerde cases wordt gemeld dat de huidige LUMI-faciliteit 'vanwege de architectuur' (en dus vermoedelijk omdat AMD-chips worden gebruikt) niet bruikbaar is. Het is in dit geval en in vergelijkbare gevallen overigens wel aannemelijk dat met substantiële inzet van machine learning engineers een versie van de gebruikte code kan worden gemaakt die wel goed werkt op AMD-chips. Voor de uitbreiding van LUMI is de inzet van NVIDIA-chips voorgenomen.

Op de langere termijn komen naar verwachting chips op de markt die gespecialiseerd zijn in AI-taken. Onder andere in China wordt gewerkt aan eigen chips voor AI-training en -inferentie. [6] Partijen als Apple, Google en Amazon ontwikkelen en gebruiken daarnaast eigen chips (zgn. TPU's – *tensor processing units* of NPU's –*neural processing units*) die met name gericht zijn op efficiënte inferentie en bijvoorbeeld specifiek zijn gemaakt voor gebruik in mobiele telefoons. [7] [8] Omdat de GPU-architectuur zeer goed aansluit bij de rekenbehoefte van hedendaagse AI en de geboden flexibiliteit van pas komt bij het trainen van modellen, komt de ontwikkeling van echt gespecialiseerde chips daarvoor echter nog niet echt op gang, en is de verwachting dat de sterke afhankelijkheid van NVIDIA zeker de komende jaren nog blijft bestaan.

Daarnaast is ook andere apparatuur nodig, zoals *netwerkapparatuur* (switches en routers) om de servers onderling en met de buitenwereld te verbinden. Een bijzonder relevant onderdeel van dit netwerk is de *interconnect*, waarmee de servers ('nodes') onderling zijn verbonden en tijdens het uitvoeren van een rekentaak razendsnel gegevens kunnen uitwisselen.

Tot slot is er grootschalige **opslagcapaciteit** nodig, met name voor de grote hoeveelheden gegevens waarmee een AI-model wordt 'getraind'. Binnen de AI-fabriek is in ieder geval een vorm van *tijdelijke* opslag nodig (deze bewaart de trainingsdata en tussenresultaten op, voor zolang een bepaalde rekentaak actief is of moet kunnen worden herstart). Deze opslagcapaciteit dient snel aanspreekbaar te zijn voor de rekencapaciteit. Daarnaast is er *permanente* opslag nodig om bijvoorbeeld de brondatasets en de getrainde modellen op te slaan. Deze permanente opslag kan buiten de AI-fabriek worden gerealiseerd (bijvoorbeeld in een eigen datacenter van de ontwikkelende organisatie) of onderdeel zijn van de fabriek (bijvoorbeeld om herbruikbare datasets op te slaan). In een scenario waarin de rekencapaciteit zich buiten Nederland bevindt kan ook worden gedacht aan een Nederlandse opslaglocatie die dient als een soort 'clearing house'. De data kunnen dan in Nederland worden opgeslagen, en voorafgaand aan training op buitenlandse infrastructuur geschoond en voorbereid. Het is ook denkbaar dat in sommige gevallen een deel van de berekeningen in Nederland kan blijven, zodat er minder of geen gevoelige data op buitenlandse infrastructuur hoeft te worden

opgeslagen. Er is hoogwaardige connectiviteit nodig tussen de permanente opslag en rekencapaciteit (als beide zich in separate datacentra bevinden is een snelle glasvezelverbinding nodig).

Merk hierbij op dat in de definitie van AI-fabriek (ook) wordt gesproken over gedistribueerde faciliteiten. Dit zou concreet betekenen dat er meerdere fysieke faciliteiten zijn (geografisch verspreid) waarover de rekenapparatuur verdeeld is, en waartussen zeer hoogwaardige connectiviteit bestaat om deze te kunnen laten samenwerken.

Op de *nodes* draait **software** die rekentaken kan ontvangen en uitvoeren. Naast een besturingssysteem is gespecialiseerde software nodig die ervoor zorgt dat de grote aantallen servers binnen een AI-fabriek kunnen samenwerken. Een laag hoger vinden we de AI-programmeeromgeving (softwarebibliotheken, bijbehorende documentatie, etc.) die ontwikkelaars in staat stellen om gebruik te maken van deze faciliteit, en algoritmes kunnen ontwikkelen die parallel worden uitgerekend over grote aantallen *nodes* tegelijkertijd. Het omzetten van een algoritme naar een *paralleliseerbaar* algoritme, dat goed werkt en geoptimaliseerd is voor een HPC-omgeving, is geen *sine cure*.

Helemaal bovenaan de lagenstapel binnen de AI-fabriek vinden we de HPC AI-rekentaken. Een **coördinator** wijst de taken toe aan de supercomputer op basis van vooraf gemaakte afspraken (tijdsverdeling, verdeling tussen soorten gebruikers, et cetera).

2.1.2 Benodigde kennis en expertise

Links in Figuur 1 worden de kennis en expertise getoond die nodig zijn bij de verschillende lagen. Deze expertise en kennis zullen in de AI-fabriek aanwezig moeten zijn. De bovenste lagen van de figuur vereisen hierbij specifieke expertise die wordt opgebouwd in interactie met de AI-faciliteit, zoals de expertise om AI-algoritmes te vertalen naar een HPC-omgeving. Dergelijke expertise vereist geen fysieke toegang tot de concrete faciliteiten. Alle beheerstaken in het digitale domein worden over het algemeen op afstand uitgevoerd via een beveiligde verbinding.

De vaardigheden die benodigd zijn bij fysieke toegang tot de fysieke faciliteiten zijn veelal generiek. Het aantal mensen dat dagelijks fysiek in de AI-fabriek zal moeten werken om de supercomputer en dataopslag aan de gang te houden, is zeer beperkt. De taken beperken zich bij dergelijke grote datacenters veelal tot het verwisselen van defecte hardware zoals servers en harde schijven, en het diagnosticeren van technische problemen [9]. Daarnaast zijn er uiteraard generieke facilitaire taken rondom het gebouw.

2.2 Inzet van een AI-fabriek

Een AI-fabriek is primair ontworpen om rekencapaciteit te bieden die kan worden ingezet voor het ontwikkelen van AI. De AI-fabriek kan echter ook worden ingezet voor

andere vormen van grootschalige data-analyse. Steeds meer data scientists raken bedreven in het werken met GPU-infrastructuur in plaats van (of in aanvulling op) CPU-infrastructuur.

In beide vormen is er een groot en voor deze analyse relevant verschil tussen de *ontwikkelfase* en de *gebruiksfase*. De bijpassende infrastructuur en ook expertise is in iedere fase wezenlijk anders.

Ontwikkelfase

In de ontwikkelfase wordt een model ontwikkeld op basis van grote hoeveelheden data. Bij bijvoorbeeld een weermodel kun je hierbij denken aan grote hoeveelheden gedetailleerde meteorologische gegevens; bij grote taalmodellen (generatieve AI) gaat het om zeer grote hoeveelheden tekst. Bij veel vormen van AI is het relevant om deze fase verder te splitsen in de *pre-trainingfase* en de *fine-tuningfase*.

Pre-trainingfase

Pre-training is de fase waarin een AI (machine learning-model) vanaf nul wordt opgebouwd. Hierbij worden grote hoeveelheden data (tekst, audio, afbeeldingen, video, etc. afhankelijk van het type model) geanalyseerd om patronen te herkennen. Deze patronen vormen het uiteindelijke generieke (ook wel *foundation*) AI-model. Voorbeelden van dergelijke (foundation) modellen zijn de *large language models* (LLM's) zoals de GPT's (*Generative Pretrained Transformer*) van OpenAI, Google Gemini of GPT-NL. Hoewel de aandacht veelal uitgaat naar grote taalmodellen zijn er ook andere soorten *foundation models*, zoals het AlphaFold-model, dat onlangs werd beloond met een Nobelprijs, waarmee proteïnestructuren kunnen worden voorspeld [10].

De pre-trainingfase vereist zowel grote hoeveelheden data (opslag- en verwerkingscapaciteit), als aanzienlijke rekenkracht om de data te analyseren en het model te creëren. Hoe meer rekenkracht, hoe sneller deze fase kan worden doorlopen. Kenmerkend is daarnaast dat er veel data moet worden uitgewisseld tussen de verschillende onderdelen van het model. Een AI-fabriek voldoet precies aan deze vereisten, en heeft daarmee een **grote meerwaarde in de pre-trainingfase**.

Fine-tuningfase

Fine-tuning is de fase waarin *foundation* modellen worden geoptimaliseerd voor bepaalde toepassingen of gebruiksscenario's. Zo kan een model worden gespecialiseerd voor bepaalde talen, typen data (bijv. juridische teksten of reisinformatie) of interacties (zoals een chat-interface). De benodigde data en rekenkracht zijn in deze fase aanzienlijk, maar beduidend minder groot dan in de pre-trainingfase. Een AI-fabriek heeft daarmee een **beperkte meerwaarde in de fine-tuningfase**.

Gebruiksfase

In deze fase passen gebruikers bestaande AI-modellen toe voor hun eigen vragen of opdrachten. Gebruikers kunnen een AI-model vragen een tekst te genereren, een

afbeelding te interpreteren of informatie te geven over een bepaald onderwerp, een weermodel kan een voorspelling maken voor de volgende dag op basis van actuele meetwaarden, et cetera.

Een AI-fabriek heeft in onze ogen **beperkte tot geen meerwaarde in de gebruiksfase**. Hoewel een HPC-faciliteit die is ontworpen voor training in principe kán worden ingezet voor toepassing (inferentie), is het voor dit doel een relatief 'dure' infrastructuur. Bij inferentie worden namelijk veel beperktere hoeveelheden data verwerkt (namelijk alleen de modelgewichten en dat wat wordt ingevoerd door de gebruiker) en veel beperktere rekenkracht per gebruiker ten opzichte van de trainingsfase. De 'zware' architectuur van de HPC-faciliteit biedt hier weinig meerwaarde boven eenvoudigere infrastructuur, die goedkoper in aanschaf en gebruik is. Voor sommige toepassingen kan het aantal gelijktijdige gebruikers uiteraard wel zeer groot zijn, maar het realiseren van de benodigde inferentiecapaciteit is dan een kwestie van het (in aantal) opschalen van relatief eenvoudige rekencapaciteit.

In het toegangsbeleid vanuit EuroHPC wordt daarnaast bepaald dat voor gebruikers vanuit de wetenschap het gebruik gericht moet zijn op het trainen, benchmarken en valideren van (foundation) modellen, en dus niet op inferentie. [11, p. 16 H3: access modes] Voor commerciële gebruikers (die aanspraak kunnen maken op 20% van de capaciteit) geldt deze vereiste niet. [11, pp. 30, par. 3.9] Desondanks lijkt het (om voornoemde reden) doelmatiger om de hiervoor beschikbare capaciteit in eerste plaats in te zetten voor pre-training en fine-tuning. Het inzetten van de AI-fabriek voor inferentie is voor commerciële gebruikers waarschijnlijk vooral een kwestie van een eventueel kostenvoordeel dat de fabriek biedt boven het alternatieve commerciële aanbod (waar mogelijk sprake is van schaarste).

2.3 Plek van een AI-fabriek in het AI-ecosysteem

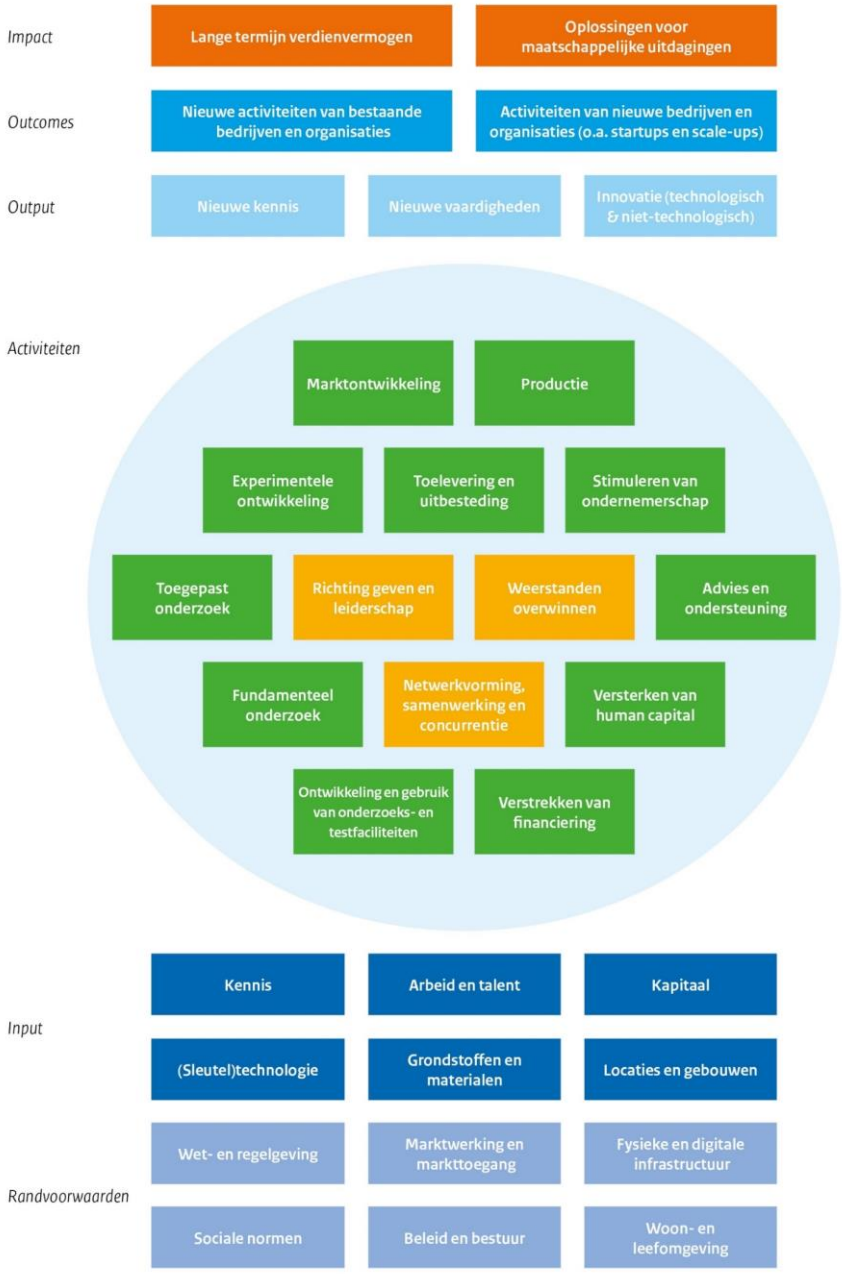
Een eventuele AI-fabriek bevindt zich niet in een vacuüm, maar is onderdeel van een breder systeem of 'ecosysteem'. Wanneer wij binnen deze verkenning spreken over het ecosysteem of AI-ecosysteem sluiten wij aan bij de definitie van een onderzoeks- en innovatie-ecosysteem zoals gehanteerd in de achtergrondstudie voor de kabinetsstrategie m.b.t. het versterken van dergelijke ecosystemen [12]:

> ***Een ecosysteem voor onderzoek en innovatie omvat een dynamische set van samenhangende actoren, activiteiten, faciliteiten en regels die van belang zijn voor het onderzoeks- en innovatievermogen van individuele actoren en groepen van actoren en, hierdoor, voor het creëren van waarde.***

Een onderzoeks- en innovatie-ecosysteem (O&I-ecosysteem) wordt gekenmerkt door verschillende vormen van inputs, throughputs, outputs, outcomes en impact, zie Figuur

2. In het midden van de figuur worden verschillende activiteiten getoond die als het ware het hart vormen van het O&I-ecosysteem.

Als we naar 'het AI-ecosysteem' in Nederland kijken zien we een opgebouwd ecosysteem met verschillende regionale hubs, waarbij de Nederlandse AI Coalitie (NL AIC) als centrale aanvoerder gezien kan worden [13]. Binnen het AI-ecosysteem wordt al jaren gewerkt aan de verschillende activiteiten zoals getoond in Figuur 2. Coördinatie van het Nederlandse AI-ecosysteem is onder meer ingevuld door de NL AIC, AINed investeringsprogramma (NGF), het Strategisch Actieplan voor AI en de regionale AI-hubs.



Figuur 2. Conceptueel kader O&I-ecosystemen. Bron: [12]

Een eventuele AI-fabriek zou binnen dit bredere AI-ecosysteem kunnen bijdragen aan verschillende activiteiten en functies, zodat het ecosysteem als geheel beter kan bijdragen aan het lange termijnverdienvermogen en oplossingen voor maatschappelijke uitdagingen (impact). De meest voor de hand liggende activiteiten zijn:

- Ontwikkeling en gebruik van onderzoeks- en testfaciliteiten
- Experimentele ontwikkeling
- Versterken van human capital
- Advies en ondersteuning
- Stimuleren van ondernemerschap

Daarnaast kan een AI-fabriek (indirect) ook een relatie hebben met andere activiteiten in het ecosysteem. Zodra er bijvoorbeeld nieuwe AI-modellen ontwikkeld worden in de AI-fabriek kan dat ook tot marktontwikkeling en productie leiden. Ook kan een AI-fabriek verband houden met de 'inputs' die binnen het ecosysteem worden ingezet, waarbij 'Kennis' en 'Arbeid en talent' binnen deze context vaak belicht worden. Een ecosysteem kan daarbij gepaard gaan met 'economische voordelen' in de vorm van **schaalvoordelen**, **scopevoordelen** en **agglomeratievoordelen**, zie het tekstkader hieronder.

Voor deze verkenning is het belangrijk om te constateren dat een eventuele AI-fabriek een onderdeel gaat uitmaken van een bredere dynamische set van samenhangende actoren, activiteiten, faciliteiten en regels. Dit maakt het complex om exact te kunnen voorspellen wat de gevolgen van een AI-fabriek zullen zijn.

Ecosysteemwerking: lessen uit de digitale hub

Eerder onderzocht Dialogic de werkingsmechanismen achter Nederland als digitaal knooppunt (ook wel: digitale hub) [14]. Hierin brachten we in kaart hoe verschillende onderdelen van de digitale infrastructuur interacteren. Mechanismes die van invloed zijn op ecosysteemwerking zijn in het kort:

- **Schaalvoordelen** is het fenomeen waarbij de productiekosten per eenheid dalen als de productie in omvang toeneemt. Dit betekent dat het voor aanbieders loont om meer aansluitingen, routes, netwerken of dienstenaanbod in eigen hand te hebben. Internetknooppunten zijn hierbij het meest concrete en meetbare voorbeeld van een activiteit met zeer sterke **netwerkeffecten**.

- **Scopevoordelen** zijn kostenvoordelen die een bedrijf kan behalen door actief te zijn in verschillende schakels van de digitale infrastructuur. Ook scopevoordelen komen veel voor tussen de schakels van de digitale infrastructuur. Partijen met eigen hyperscale datacentra kiezen er bijvoorbeeld steeds vaker voor om internationale datakabels over land of zee aan te leggen tussen eigen

datacentra en naar andere datacenterhubs. Bij aanbieders van mobiele netwerken valt op dat zij eveneens een sterke positie in de vaste markt hebben verworven.

- **Agglomeratievoordelen** treden op als gevolg van de fysieke nabijheid van partijen en activiteiten. Agglomeratievoordelen zijn sterk aanwezig bij het samenspel tussen datakabels, datacentra, internetknooppunten, peering, hosting en cloud. Lokale aanwezigheid van deze aanbieders versterkt elkaar. Cloudaanbieders die hoogwaardige softwarediensten ontwikkelen, gebruiken datacentra om de benodigde rekenkracht en dataopslag aan klanten aan te bieden. Voor grote datacentra is vestiging in de buurt van internationale datakabels en andere datacentra interessant. Zo kunnen zij een groter gebied bedienen. Op hun beurt is het voor zeekabels weer interessant om aan te sluiten op internetknooppunten. Daar komen verschillende wereldwijd opererende netwerken bij elkaar.

Doorwerking van de digitale infrastructuur in de rest van de economie

De afhankelijkheid van de digitale infrastructuur is, zoals valt te verwachten, het grootst in de digitale economie. Daar bieden partijen hun economische activiteiten volledig aan door middel van digitale technologie. We kunnen hierbij denken aan e-commerce, IT-bedrijven of online platforms. Ook bedrijfstakken buiten de digitale economie zijn inmiddels echter sterk afhankelijk van de digitale infrastructuur.

Voor bepaalde economische activiteiten zien we dat heel specifieke, hoogwaardige digitale infrastructuur een doorslaggevende vestigingsfactor is. Dit zijn bijvoorbeeld bedrijven die diensten leveren rondom cloud, datacentra, hosting en internetknooppunten, of bedrijven die sterk van deze diensten afhankelijk zijn, zoals ‘cloud gaming’. Voor de meeste sectoren is de digitale infrastructuur vooral een belangrijke factor in een mix met andere belangrijke vestigingsfactoren.

De vergelijking met AI-fabriek / AI-faciliteit

Wanneer we de AI-fabriek als specifieke vorm van digitale (onderzoeks)infrastructuur beschouwen, dan kunnen wij ons goed voorstellen dat voorgaande schaal-, scope- en agglomeratievoordelen op gang gaan komen. Ook kan de faciliteit zelf baat hebben bij het bestaande sterke ecosysteem rondom de digitale infrastructuur. We merken hierbij op dat deze voordelen dus (veel) verder gaan dan simpelweg de technische voordelen van de fysieke nabijheid, maar juist ook betrekking hebben zaken als de toegang tot kennis en ervaring, gemak van zakendoen en de digitale vaardigheden van afnemers.

3 Meerwaarde van een AI-fabriek

In dit hoofdstuk geven we een bondige samenvatting van de verschillende cases waarin de argumentatie is uitgewerkt voor de meerwaarde van een AI-faciliteit. Deze cases betreffen respectievelijk de meerwaarde voor de wetenschap (paragraaf 3.1), de publieke sector (paragraaf 3.2) en het bedrijfsleven (paragraaf 3.3). Dit hoofdstuk betreft dus een synthese, nog geen analyse (dat volgt in het volgend hoofdstuk).

3.1 Meerwaarde voor de wetenschap

De verwachte meerwaarde voor de wetenschap van een AI-fabriek is groot (zie o.m. [15]). Waar AI-onderzoek de afgelopen jaren veelal op zichzelf stond, is AI in toenemende mate van belang voor andere vakgebieden. Denk hierbij aan het AlphaFold-model, een AI-model dat onlangs werd beloond met een Nobelprijs voor Chemie. Deze spill-over heeft mogelijke grote impact, en leidt tot toenemende behoefte aan rekenkracht (om meer projecten met grotere datasets te ondersteunen) en ondersteuning (voor AI-toepassingen in niet-technologische vakgebieden). Ten behoeve van deze verkenning is een wetenschapscasus geschreven. Daarnaast hebben NWO en SURF in afstemming met de NWO adviescommissie Digitalisering Onderzoek een studie gedaan naar de wetenschappelijke behoeftes aan rekenkracht in het algemeen (ook niet-AI) [16].

Factoren die van invloed zijn op de verdere versterking van (AI binnen) wetenschapsgebieden zijn:

- **Voldoende rekenkracht**. Hierbij wordt opgemerkt dat de bestaande nationale supercomputer van SURF (de Snellius) al tegen haar grenzen loopt en (zeker naar de toekomst) onvoldoende rekenkracht biedt voor state-of-the-art modellen en het toenemend aantal aanvragen.
- **Voldoende expertise** t.a.v. wet- en regelgeving rondom AI en de grote hoeveelheid beschikbare AI-modellen en de snelle ontwikkeling van geavanceerde modellen is essentieel om wetenschappers te ondersteunen.
- **Beschikbaarheid data** voor de ontwikkeling van AI-modellen. Verschillende vakgebieden hebben hierbij te maken met sensitieve data (voornamelijk in de sociale en medische wetenschappen), waarvoor beperkingen gelden voor hoe en waar deze verwerkt kunnen worden.
- **Positionering van Nederland als hightech land**. Aantrekken en behouden van (internationaal) talent (onderzoekers en studenten), samenwerking met het bedrijfsleven, en het aantrekken van onderzoeksbeurzen worden deels bepaald door het imago van de Nederlandse wetenschap.

Vanuit de wetenschap lijkt er samengevat behoefte aan een AI-fabriek.

3.2 Meerwaarde voor de publieke sector

Binnen overheidsdiensten wordt steeds vaker AI ingezet voor de uitvoering van taken [17]. AI brengt kansen voor een efficiënter bedrijfsvoering en dienstverlening, maar moet verantwoord ingezet worden. Ten behoeve van deze verkenning is door BZK een publieke sector case geschreven op basis van gesprekken met vertegenwoordigers uit verschillende overheidsdiensten. De uitkomsten van deze gesprekken komen gedeeltelijk overeen met recente studies over het gebruik van AI door de overheid.[4] Gesteld moet worden dat de mate waarin een overheidsorganisatie behoefte heeft aan meer AI-rekenkracht, sterk verschilt per organisatie. Factoren die invloed hebben op deze behoefte zijn:

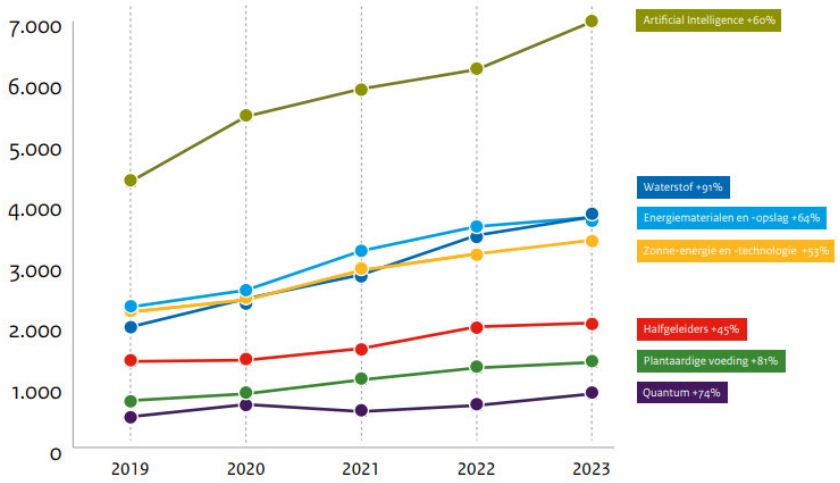
- **Data-delen.** Organisaties hebben de mogelijkheid om data te delen binnen een AI-fabriek, maar er zijn zorgen over vertrouwelijkheid en de bereidheid tot delen van gevoelige gegevens.
- **Beschikbare AI-expertise.** Aangezien sommige organisaties de benodigde expertise missen, zal een AI-fabriek ook een advies en begeleiding moeten bieden naast de beschikbare rekenkracht.
- **Mogelijkheden voor toegang.** Snelle en directe toegang, zonder complexe aanvraagprocedures, wordt genoemd als belangrijk (vooral voor kleine projecten).
- **Complexiteit van AI-model en -toepassing.** Slechts een beperkt aantal organisaties is nu nog in staat grote, complexe modellen te ontwikkelen. De ambities variëren van kleinschalige toepassingen tot grotere, geavanceerde AI-projecten.
- **Beveiliging van data**. Voor sommigen volstaat beveiliging die in lijn is met Europese wet- en regelgeving; anderen zijn terughoudend om data buiten eigen muren te delen.

Voor de publieke sector lijkt er dus geen noodzaak voor een AI-fabriek, maar voorziet een AI-fabriek wel in een aantal behoeften. Deze variëren van testen op bias, verder onderzoek doen naar generatieve AI, schaarse expertise en rekenkracht bundelen en afspraken maken over datagebruik/samenwerking. Bij verschillende overheidsdiensten bestaat de overtuiging dat (de inzet van) AI steeds belangrijker kan worden en deze behoeften in de toekomst urgenter worden.

[4] In de public sector case zijn o.a. de bevindingen uit verkennend onderzoek van TNO [37] en een Europees onderzoek naar AI-adoptie (EU study calls for strategic AI adoption to transform public sector services | Shaping Europe's digital future (europa.eu)) meegenomen.

3.3 Meerwaarde voor het bedrijfsleven

Op het gebied van R&D bij bedrijven speelt AI een grote rol. Dit is onder meer zichtbaar in het grote en sterk stijgende aandeel in arbeidsjaren voor de WBSO-regeling op het subthema AI, zie Figuur 3. Voor deze verkenning heeft AiNed een bedrijfscasus uitgewerkt en heeft het adviesbureau KplusV in opdracht van NOM/Provincie Groningen een impact-analyse gedaan van een AI-fabriek in Groningen [18].



Figuur 3. Aantal WBSO-arbeidsjaren in 2023 voor belangrijkste subthema's. Bron: [19]

Om de Nederlandse economie internationaal concurrerend te houden, is het van belang dat er voldoende mogelijkheden zijn voor nieuwe bedrijven (startups) om zich te onderscheiden en voor bestaande bedrijven om nieuwe activiteiten te ontplooien. Factoren die van invloed zijn op de verdere versterking van het internationaal concurrerend vermogen van de Nederlandse economie zijn:

- **Voldoende rekenkracht**. Het is voor veel bedrijven nog niet rendabel om eigen AI-faciliteiten op te bouwen, zeker wanneer AI niet de kernactiviteit is. Voor (een klein gedeelte van de) startups geldt dat veelal zij nog niet de middelen hebben om voldoende rekenkracht op te bouwen en dit dus een drempel vormt voor hun doorontwikkeling.
- **Voldoende expertise**. Veel bedrijven missen de benodigde expertise. Door schaarste op de arbeidsmarkt is het ook niet altijd goed mogelijk om deze binnen bedrijven op te bouwen.
- **Beschikbaarheid data** voor de ontwikkeling van AI-modellen. Er is voor bedrijven en voornamelijk startups behoefte aan publieke data om AI-modellen op te trainen en testen.
- **Positionering van Nederland als hightech land**. Aantrekken en behouden van (internationaal) talent en bedrijven (startups, mkb en grootbedrijven) wordt deels bepaald door het imago van het Nederlandse vestigingsklimaat, inclusief de aanwezigheid van toponderzoekers.

Vanuit het bedrijfsleven lijkt er samengevat behoefte aan een AI-faciliteit.

4 Meta-analyse

In dit hoofdstuk gaan we in op de analyse die we hebben uitgevoerd van de onderbouwing van de meerwaarde van een AI-fabriek (vanuit de drie in het vorige hoofdstuk beschreven perspectieven) in relatie tot de beleidsopties voor het al dan niet realiseren ervan.

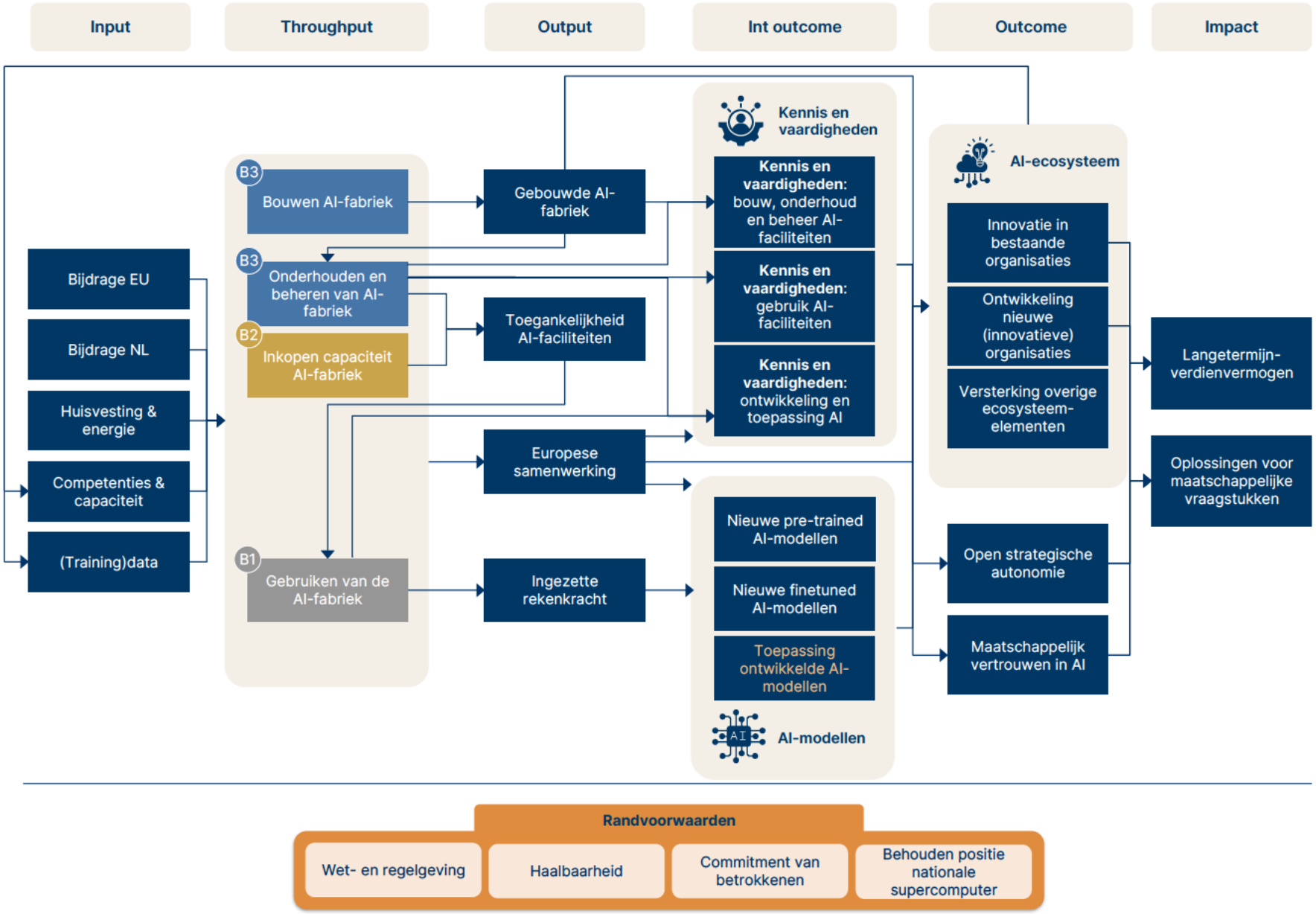
4.1 Analysekader

De meta-analyse beoogt systematisch uit te werken welke consequenties verwacht kunnen worden van de drie beleidsopties. We gebruiken hiervoor een *theory of change* (ToC); dit is een uitgebreide beschrijving die de veronderstellingen uitlegt hoe en waarom een gewenste verandering kan worden verwacht. In dit verband wordt gekeken naar de verschillen in de (lange termijn) doelen die de drie beleidsopties dienen en welke effecten het heeft op het AI-ecosysteem in Nederland. Het is belangrijk dat hierbij de randvoorwaarden worden nageleefd en dat de kosten in verhouding blijven tot de gewenste uitkomsten. In dit kader worden de volgende elementen onderzocht:

- ***Input:*** de financiële middelen, mensen, kennis en andere middelen die nodig zijn voor het uitvoeren van de geplande activiteiten. Zonder deze input kunnen de gewenste activiteiten en veranderingen niet worden gerealiseerd.
- ***Throughput:*** de specifieke acties die worden uitgevoerd binnen de beleidsopties.
- ***Output*:** de directe resultaten van de beleidsopties, zoals de gebouwde faciliteiten of de diensten die kunnen worden geleverd.
- ***Intermediate outcome:*** de specifieke korte- tot middellange termijn veranderingen die worden gerealiseerd, volgend uit de resultaten van de beleidsopties.
- ***Outcome:*** de specifieke lange-termijn veranderingen die worden gerealiseerd binnen het AI-ecosysteem en de maatschappij als gevolg van de implementatie van de beleidsopties.
- ***Impact:*** de lange termijn veranderingen en verwachtingen die optreden als gevolg van het uitvoeren van de beleidsopties. Over het algemeen is de impact niet meetbaar of anderszins herleidbaar tot de acties, maar geeft dit niveau de context weer *waarom* doelstellingen worden nagestreefd.
- ***Randvoorwaarden:*** noodzakelijke voorwaarden die vervuld moeten zijn om de beleidsopties uit te kunnen voeren. Als niet aan deze randvoorwaarden wordt voldaan, kan de verandering niet op de gewenste manier plaatsvinden.

De meta-analyse voeren we uit door voor elk onderdeel van de ToC de drie beleidsopties te vergelijken op basis van de uitgewerkte cases en ander aangeleverd materiaal. De ToC is weergeven in Figuur 4.

Figuur 4. Theory of change voor een AI-fabriek



4.2 Input

Bijdrage EU

Binnen de EuroHPC JU-calls kan geld worden aangevraagd voor het ontwikkelen van een hub en voor het bouwen van een AI-fabriek. In beide gevallen geldt dat EuroHPC JU de Nederlandse inleg zal verdubbelen. Binnen dit criterium beschouwen we de mogelijkheid om aanspraak te maken op Europese gelden als een voordeel op zichzelf.

Met B3 is de Nederlandse inbreng het grootst, en daarmee ook de Europese bijdrage aan Nederland. Met B2 geldt deze 'verdubbelaar' alleen over de middelen voor de hub in Nederland (niet over de bijdrage aan het internationale consortium, waarvoor de Europese bijdrage aan het hostland wordt overgemaakt). Met B1 ontvangt Nederland geen middelen van de EU.

De EuroHPC JU-call is in concurrentie tussen lidstaten. Er bestaat dus een risico dat aangevraagde gelden niet worden toegekend, wanneer andere lidstaten betere voorstellen indienen. Dit risico kunnen we niet goed wegen.

Tabel 1. Ranking beleidsopties op het criterium Bijdrage EU

	3e (Slechtste)	2e	1e (Beste)
Bijdrage EU	B1	B2	B3

Bijdrage NL

De investeringen in een AI-fabriek zijn substantieel. Dit omvat de benodigde financiële middelen voor de concrete faciliteiten (zoals de grote hoeveelheid GPU's), de salariskosten van het onderhoudend personeel en de expertise in de hub, en de energiekosten (de energiekosten werken we nader uit onder *Huisvesting & energie*).

Simpel gesteld lopen de scenario's op in prijs. Dit uit zich echter niet alleen in de concrete euro's. Hoewel er nu wordt uitgegaan van een investeringsbedrag wat met incidentele middelen kan worden gedekt, wordt vanuit de wetenschap gevraagd om een duurzaam financieringsmodel [16]. Een AI-fabriek (en andere HPC-faciliteiten, zoals Snellius) moet worden afgeschreven en op termijn vervangen. Als gebruik van de te bouwen AI-fabriek niet meer financieel rendabel is, moet deze worden afgebroken, gerecycled en vervangen. Deze punten spelen voornamelijk voor B3. Daarbij speelt met B3 dat er een vrij groot financieel risico wordt genomen dat gedekt moet worden; SURF (als beoogde aanvrager en penvoerder voor een Nederlandse AI-fabriek) kan dit financieel risico niet zelfstandig dragen [20]. Deze risico's zijn bij B1 uiteraard niet aanwezig,

maar ook bij B2 vele malen kleiner doordat de beheerslasten primair bij een Europese partner liggen.

Tabel 2. Ranking beleidsopties op het criterium Bijdrage NL

	3e (Slechtste)	2e	1e (Beste)
Bijdrage NL	B3	B2	B1

Huisvesting & energie

De concrete faciliteiten en de mensen met expertise dienen gehuisvest te worden in een (of meerdere) gebouw(en). Voor de locatie van de huisvesting van de concrete faciliteiten speelt daarbij bovendien de beschikbaarheid van (duurzame) energie.

B1 doet geen additioneel beroep op huisvesting en energie. B2 scoort gelijk aan B1, omdat bijvoorbeeld LUMI al beschikt over huisvesting en van groene stroom wordt voorzien. Naar verwachting wordt de LUMI AI-fabriek CO_2-negatief (waarmee het dus mogelijk zelfs beter scoort dan scenario 1) [21, p. 5].

B3 geeft de grootste uitdagingen op huisvesting en vooral energie. Voor huisvesting bestaan mogelijkheden, onder meer een leegstaand pand in Groningen. Dit pand heeft een energieaansluiting en is dicht bij een toekomstig windpark [20]. Toch zal de AI-fabriek vragen om een grote hoeveelheid (groene) stroom in een nationale energiemarkt die al grote uitdagingen heeft om alle bedrijven van voldoende stroom te voorzien. Toekomstige ontwikkelingen in de Nederlandse stroomprijs vormen een risico voor financiële houdbaarheid van het draaien van de AI-fabriek [16], [20].

Tabel 3. Ranking beleidsopties op het criterium Huisvesting en energie

	3e (Slechtste)	2e	1e (Beste)
Huisvesting en energie	B3		B2 B3

Competenties & capaciteit

Voor het bouwen, beheren en gebruiken van een AI-fabriek zijn mensen nodig met de juiste competenties. Deze zijn nodig 'aan de voorkant' (zonder mensen met competenties kan je een AI-fabriek niet bouwen en gebruiken), maar ontstaan ook 'aan de achterkant' (een AI-fabriek heeft een mogelijk positief effect op het vergroten van de

groep mensen met de juiste competenties; dit is weergeven in Figuur 4 met een pijl die terugkeert van de outcome *AI-ecosysteem* naar de input *Competenties & capaciteit*).

Nederland beschikt over wetenschappers en bedrijven die onderzoek doen naar en met AI. In zowel de wetenschaps- als de bedrijfscasus wordt het belang genoemd om dit talent voor Nederland te behouden, naast het opbouwen van nieuw talent. B3 heeft hier mogelijk de meeste toegevoegde waarde, maar geeft (in ieder geval op korte termijn) het risico dat de AI-fabriek concurreert met kennisinstellingen en bedrijven om schaars AI-talent. Op langere termijn geeft B3 mogelijk schaalvoordelen door bundeling van expertise en verbeterde samenwerking met kennisinstellingen en het bedrijfsleven. Dit risico op de korte termijn is bij B2 minder groot, omdat wordt samengewerkt met experts uit het internationale consortium. B2 en B3 scoren daarom gelijk op dit criterium.

B1 vereist geen actieve opbouw van competenties & capaciteit, maar geeft dus ook op lange termijn geen positief effect hierop.

Tabel 4. Ranking beleidsopties op het criterium Competenties & capaciteit

	3e (Slechtste)	2e	1e (Beste)
Competenties en capaciteit	B1		B2, B3

(Training)data

Het trainen en toepassen van AI-modellen is zeer sterk afhankelijk van de beschikbare (training)data. Hoe meer data, en hoe hoger de kwaliteit hiervan, hoe beter de mogelijke AI-modellen.

In de verschillende cases wordt onderscheid gemaakt tussen de consequenties voor (1) gebruikers om hun data naar de AI-fabriek te brengen ter analyse, en (2) de mogelijkheid om een centrale publieke dataset aan te bieden om (nieuwe) AI-modellen op te trainen en testen.

Een AI-fabriek binnen Nederland (B3) maakt het mogelijk om gevoelige data binnen de AI-fabriek te analyseren, die niet (zomaar) bij commerciële partijen (denk aan AI-infrastructuur van Google) of in het buitenland (zoals LUMI in Finland) kan worden verwerkt. Gevoelige data is in alle cases aanwezig, zoals sociale en medisch wetenschappers, overheidsdiensten met data over burgers, en bedrijven met concurrentiegevoelige data. Hoewel het juridisch waarschijnlijk strikt genomen mogelijk is om data in dergelijke gevallen ook op een Europese supercomputer te verwerken (met voldoende cybersecuritymaatregelen), verwachten gesprekspartners dat dit een veel moeizamer

proces zal zijn. Een AI-fabriek binnen Nederland zal alsnog niet altijd voldoen: in alle cases zijn er partijen wiens data dermate gevoelig is, dat deze niet buiten de muren van de eigen instelling mag worden verwerkt. Het zal echter meer *use cases* mogelijk maken, en biedt ook meer perspectief in onderhandelingen met rechthebbenden om op een later moment data alsnog buiten de muren, maar binnen de grenzen te verwerken. Met B1 zal dit zelfs nog moeilijker blijken, door afhankelijkheid van commerciële infrastructuur en het gebrek aan invloed van dataverwerking binnen AI-fabrieken.

Daarnaast maakt B3 een fysieke faciliteit met dataopslag mogelijk om binnen Nederland te werken aan een publieke dataset waarop AI-modellen kunnen worden getraind, gefinetuned of getest. Een dergelijke dataset is in principe ook mogelijk bij B1 en B2, maar vergt dan gebruik van een ander datacentrum dan de AI-Fabriek. Overigens is het geen vaststaand gegeven dat een AI-fabriek binnen Nederland zal leiden tot een dergelijke publieke dataset.

Tabel 5. Ranking beleidsopties op het criterium (Training)data

	3e (Slechtste)	2e	1e (Beste)
(Training)data	B1	B2	B3

4.3 Throughput

De throughput omvat de beleidsopties, die zich als volgt tot elkaar verhouden:

- **B1**: een AI-fabriek kan enkel worden gebruikt via competitieve EuroHPC calls.
- **B2**: voegt aan B1 toe dat Nederland capaciteit 'inkoopt' bij een AI-fabriek, en er een lokale 'hub' (bijvoorbeeld een kenniscluster, al dan niet met een fysieke faciliteit) wordt gerealiseerd die de toegang tot deze capaciteit faciliteert.
- **B3**: voegt aan B1 en B2 toe dat een eigen AI-fabriek wordt gebouwd en onderhouden, de ingekochte capaciteit AI-fabriek bevindt zich dan op Nederlandse bodem.

4.4 Output

Gebouwde AI-fabriek

Met B2 en B3 investeert de Nederlandse overheid in de bouw van een AI-fabriek. Bij B1 komen er alsnog AI-fabrieken in andere Europese lidstaten, die ook toegankelijk zullen worden voor wetenschappers, bedrijven en overheden in Nederland. Om deze reden is er geen 0-scenario, aangezien er sowieso AI-faciliteiten beschikbaar komen op

Europees niveau. De bouw van een AI-fabriek is geen doel op zichzelf, om deze reden geven we voor dit onderdeel geen ranking. Het is strikt geen criterium, maar is opgenomen in het analysekader voor de volledigheid.

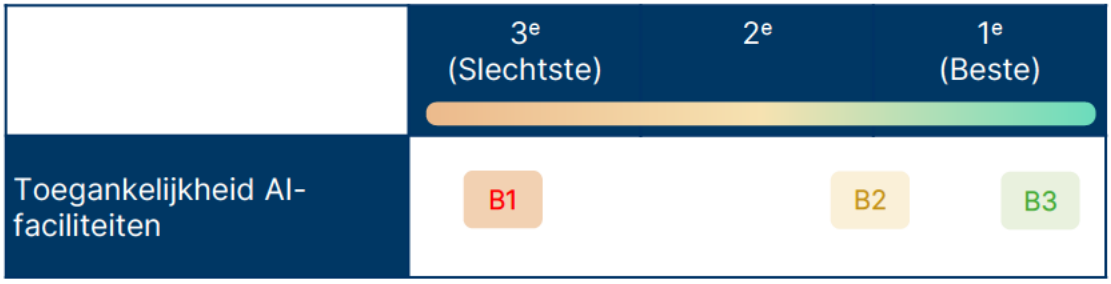
Toegankelijkheid AI-faciliteiten

Onder toegankelijkheid van de AI-faciliteiten verstaan we de mate waarin Nederlandse gebruikers (wetenschappers, bedrijven, overheden) gemakkelijk toegang krijgen tot een AI-fabriek.

Wanneer toegang krijgen tot een AI-fabriek een lange en moeizame doorlooptijd kent, haken gebruikers af. Bij huidige HPC-infrastructuur is gebleken dat drempels om gebruik te maken van rekenkracht gebruikers weerhoudt. Zo geven startups aan dat zij geen gebruik maken van EuroHPC doordat zij bureaucratische processen en lange wachttijden vrezen [22]. Voor de huidige nationale supercomputer zijn wachttijden ingekort doordat wetenschappers kleine aanvragen direct bij SURF (als beheerder van de supercomputer) kunnen doen en instellingen tijd kunnen reserveren op de supercomputer ten behoeve van hun medewerkers [23].

Met B3 wordt de toegankelijkheid van de AI-fabriek het sterkst mogelijk gemaakt. Een Nederlandse beheerder kan zelf aanvragen prioriteren en rekenkracht toewijzen binnen het Nederlandse deel. Ook kunnen minder technische gebruikers worden ondersteund. Met B2 kan deze ondersteuning ook worden geleverd vanuit de hub. Nederland krijgt met B2, afhankelijk van de gemaakte afspraken, zeggingskracht over een gedeelte van een Europese AI-fabriek, maar de snelheid waarmee Nederlandse gebruikers toegang krijgen is afhankelijk van gemaakte afspraken. De mate waarin B2 minder goed scoort dan B3 is dus enigszins onzeker en afhankelijk van de afspraken die gemaakt moeten worden binnen B2. Met B1 wordt geheel niet gewerkt aan toegankelijkheid van AI-faciliteiten, maar moeten Nederlandse gebruikers in competitie rekenkracht aanvragen bij EuroHPC JU; dit vertaalt zich naar verwachting naar langere wachttijden en hogere kans op afwijzingen.

Tabel 6. Ranking beleidsopties op het criterium Toegankelijkheid AI-faciliteiten



Europese samenwerking

Europese samenwerking is van belang voor de Nederlandse wetenschap en economie om internationaal concurrerend te zijn en blijven (zie ook [24]).

B2 stimuleert Europese samenwerking het sterkst, doordat Nederland een stevige positie inneemt in het LUMI-consortium dat tot op heden bestaat uit Finland, België, Tsjechië, Denemarken, Estland, IJsland, Noorwegen, Polen, Zweden en Zwitserland [25].

B3 resulteert waarschijnlijk in minder Europese samenwerking op de AI-fabriek, omdat de meeste potentiële Europese partners zich al verbonden hebben aan andere initiatieven (waaronder LUMI). Een Nederlands consortium ligt dan het meest voor de hand [20]. Daarentegen kan B3 de Europese of internationale positie van wetenschappers en bedrijven verstevigen en onderscheidend vermogen geven, wat op langere termijn kan leiden tot meer Europese samenwerking op R&D-projecten. Ook komt het gedeelte van de capaciteit dat wordt gefinancierd door EuroHPC JU ter beschikking voor Europese aanvragers, wat samenwerking met de experts in Nederland stimuleert.

B1 stelt Europese samenwerking op achterstand. Nederland doet niet mee aan Europese consortia die wel investeren in een AI-fabriek, en heeft minder onderscheidend vermogen voor latere R&D-projecten waarin AI een rol speelt.

Tabel 7. Ranking beleidsopties op het criterium Europese samenwerking

	3e (Slechtste)	2e	1e (Beste)
Europese samenwerking	B1	B3	B2

Ingezette rekenkracht

De uiteindelijke hoeveelheid rekenkracht die beschikbaar komt voor (Nederlandse) gebruikers verschilt substantieel tussen B1 enerzijds en B2 en B3 anderzijds. Met B2 en B3 investeert Nederland immers substantieel in (toegang tot) rekencapaciteit, terwijl met B1 alleen de bestaande middelen worden ingezet.

In de verschillende cases wordt benadrukt dat de nu beschikbare rekencapaciteit (in de vorm van Snellius) volstrekt ontoereikend is voor (met name) het *pre-trainen* van *foundation* modellen. Ter illustratie van de 'achterstand': de recent toegevoegde GPU-rekencapaciteit van Snellius is al voor ongeveer een jaar gereserveerd voor het trainen van GPT-NL, een model dat zich naar verwachting qua prestaties en omvang zal bevinden op het niveau van GPT-3.5 (en dat model bestaat sinds maart 2022; inmiddels zijn reeds drie nieuwe versies uitgebracht). De voorgestelde schaalvergroting met B2 en B3 maakt een substantiële sprong in de maximale omvang van de modellen mogelijk.

Hoewel de 'achterstand' in termen van tijd zeer zichtbaar is wanneer we kijken naar het pre-trainen van grote taalmodellen, blijft de hamvraag (wanneer Nederland

substantieel zou investeren) in hoeverre de behoefte voor het pre-trainen van *foundation* modellen (waaronder nadrukkelijk ook niet-taalmodellen) in de toekomst blijft bestaan en niet 'afvlakt'. Het is met de huidige kennis van AI niet mogelijk hier een sluitend antwoord op te geven. We stellen wel vast dat er (zeker vanuit de wetenschap) zeer waarschijnlijk hoe dan ook wel een nuttige invulling kan worden gevonden voor grote hoeveelheden rekenkracht. De combinatie van lage kosten en grote hoeveelheid rekenkracht lijkt daarnaast voor Nederlandse start-ups een nuttige behoefte in te vullen.

In de geanalyseerde cases komt naar voren dat er naast het *pre-trainen* van *foundation* modellen ook (veel) behoefte is aan het *finetunen* van bestaande modellen en uiteraard aan het gebruik van de modellen (*inference*). De cases geven onvoldoende onderbouwing voor de verdeling van de totale behoefte over deze categorieën. Dit is echter wel zeer relevant, omdat de AI-fabriek vooral een substantiële toevoeging is als het gaat om *pre-training*, in mindere mate aan *finetuning*, en niet direct veel toevoegt als het gaat om *inference*. Wat de beoordeling lastig maakt is dat de behoefte aan rekencapaciteit in een aantal gevallen aanbodgedreven zou kunnen zijn (*"build it and they will come"*). Desondanks verwachten we vooral vanuit de wetenschap en start-ups en mogelijk grotere bedrijven een substantiële behoefte op het gebied van *pre-training*. Publieke organisaties en het mkb zijn, is onze verwachting, in eerste instantie op zoek naar *inference*-capaciteit en (voor de meer vergevorderde toepassingen) naar *fine-tuning*, waar een AI-fabriek zich minder goed voor leent.

In de wetenschapscasus wordt tot slot beargumenteerd dat het nuttig is om studenten toegang te geven tot infrastructuur om AI-modellen te trainen (*"een GPU voor iedere student").* Eenzelfde argument wordt in de bedrijfscasus met *"een GPU voor data scientist van een AI-gerelateerde startup"*. Hoewel dit al snel tot een behoorlijke capaciteitsbehoefte leidt gezien de aantallen studenten en data scientists (en daarmee de investering zou kunnen rechtvaardigen), zien we dit niet als een overtuigend argument voor het investeren in HPC-capaciteit. Allereerst kun je je afvragen of de capaciteit voor individuen gelijktijdig nodig is en of het individuen lukt de volledige capaciteit te benutten. Belangrijker nog is dat het gaat om een heel ander soort behoefte. Kenmerkend aan een HPC-faciliteit is namelijk dat dit (vanwege de enorme schaal, de extreem snelle interconnect, de beschikbare opslagcapaciteit, etc.) veel hoogwaardigere capaciteit biedt dan simpelweg een verzameling van individueel inzetbare GPU's. Het inzetten van een dergelijke capaciteit voor een groot aantal kleine 'jobs' is mogelijk maar (tenzij het gaat om ongebruikte capaciteit) niet erg doelmatig.

Wanneer we kijken naar de *locatie* van de AI-fabriek (het verschil tussen B2 en B3), dan stellen we vast dat er *in theorie* nauwelijks verschillen zijn als het gaat om de ingezette rekenkracht: het moet (afgezien van beperkingen van het 'buiten de deur' c.q. 'buiten Nederland' verwerken van data) goed mogelijk zijn om rekencapaciteit in te zetten, of deze zich nu in Nederland of Finland bevindt. Er spelen wellicht enkele praktische knelpunten (zoals het kunnen overbrengen van grote hoeveelheden data

van en naar de AI-fabriek), maar hierin zien we geen sterke argumenten voor een keuze tussen beide beleidsopties. Het zelf realiseren van een AI-fabriek biedt wellicht meer controle over diverse implementatiekeuzes (zoals architectuur), maar afgezien van de keuze voor het merk GPU of bijvoorbeeld de verhouding tussen CPU- en GPU-capaciteit, lijken die keuzes uiteindelijk niet een significant voordeel voor Nederland op te leveren, gelet op de (huidige) homogeniteit van de implementatie van AI-modellen.

Met B3 komt meer tijd beschikbaar van de ingezette rekenkracht, doordat Nederland zelf de rekenkracht beheert (m.u.v. het gedeelte dat wordt betaald door EuroHPC JU). B3 biedt om die reden een voordeel ten opzichte van B2. Andersom verwachten we dat de maximum rekencapaciteit per euro die beschikbaar is met B2 groter is dan met B3, waardoor die optie juist wat aantrekkelijker is. We concluderen daarom dat beleidsopties 2 en 3 (in ieder geval voor het doel van deze vergelijking) gelijkwaardig zijn.

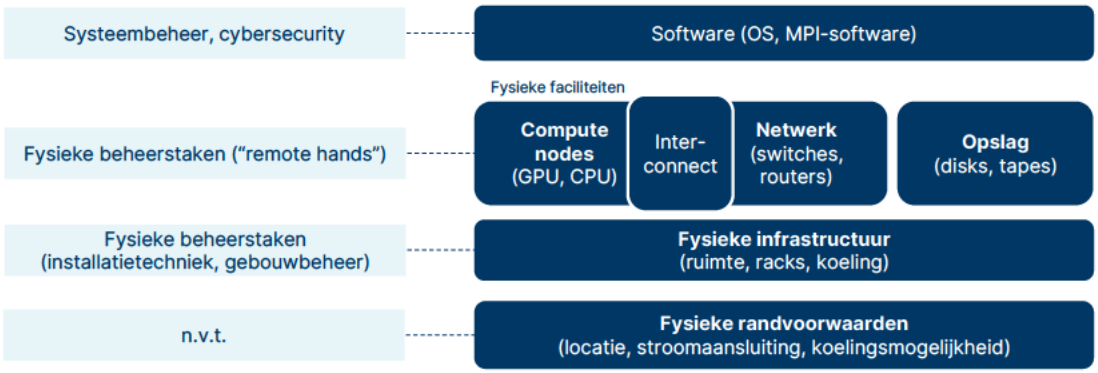
4.5 Intermediate outcome

4.5.1 Kennis en vaardigheden

Een belangrijk effect van de beleidsopties is de mate waarin Nederland kennis en vaardigheden opbouwt ten aanzien van AI. We onderscheiden drie vormen van kennis en ervaring, die we hieronder uitwerken.

Bouw, onderhoud en beheer AI-faciliteiten

Met B3 wordt met de start van de nationale AI-fabriek de meeste ervaring opgedaan, allereerst met de bouw, en vervolgens het onderhoud en beheer (zie onderstaande figuur met uitsnede van relevante lagen uit het eerdere Figuur 1).



Figuur 5. Relevante lagen van een AI-fabriek voor kennis over de bouw, onderhoud en beheer van AI-faciliteiten

T.a.v. de bouw is de mate waarin kennis en vaardigheden worden opgedaan, afhankelijk van of de uitvoering door een Nederlandse partij wordt gedaan. Hier kan gericht op aangestuurd worden.[5]

Met B2 wordt er beperkt kennis/vaardigheden opgedaan op het vlak van de bouw van de fabriek. Het feit dat je als land aan tafel zit bij deelname van toekomstige Europese AI-fabrieken, zou mogelijk kunnen resulteren in bepaalde mate van kennis over bouw (bijvoorbeeld wat de afwegingen zijn die bij een dergelijk project worden gemaakt). Expertise t.a.v. onderhoud en beheer zit ook hoofdzakelijk bij de fysieke faciliteit, dus is de invulling van de Nederlandse hub bepalend voor de mate waarin je hier kennis en vaardigheden op kunt ontwikkelen met B2.

Met B1 wordt er niet geïnvesteerd in de bouw van een AI-fabriek en ook niet in een hub. De beschikbare kennis en vaardigheden blijven daarmee op hetzelfde niveau (en gelieerd aan Snellius).

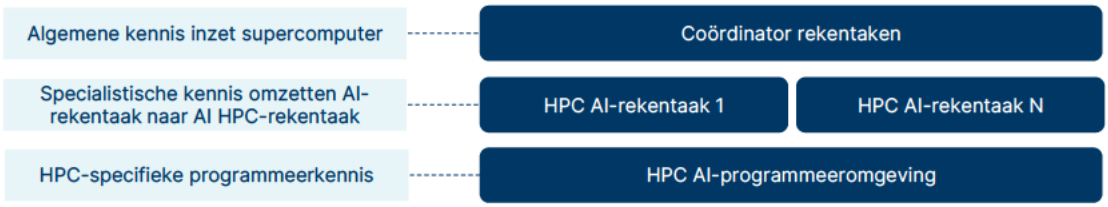
Tabel 8. Ranking beleidsopties op het criterium Kennis & vaardigheden: Bouw, onderhoud en beheer AI-faciliteiten

	3e (Slechtste)	2e	1e (Beste)
Bouw, onderhoud en beheer AI-faciliteiten	B1	B2	B3

[5] Dit betreft enkel voor de kennisontwikkeling. Wat betreft lokaal landen van de financiële investeringen geldt dat van de totale investering van B3 slechts zo’n 10% lokaal landt, doordat de meeste onderdelen van een AI-fabriek niet binnen Nederland ontwikkeld (kunnen) worden [18].

Gebruik AI-faciliteiten

Met kennis en vaardigheden t.a.v. de gebruik van AI-faciliteiten doelen we op de expertise die wordt verkregen door praktisch aan de slag te gaan met faciliteit zelf. Denk aan de onderstaande lagen uit Figuur 1.



Figuur 6. Relevante lagen van een AI-fabriek voor kennis over gebruik van AI-faciliteiten

De bepalende factor voor deze dimensie is de toegankelijkheid van de faciliteit. Met B3 is de toegankelijkheid van de AI-faciliteit waarschijnlijk het grootst, zie eerder *Toegankelijkheid AI-faciliteiten*. Een omvangrijke AI-faciliteit in Nederland biedt de grootste mogelijkheden voor Nederlands AI-talent om gebruik te maken van de faciliteit. Een onzekere factor daarbij is echter of de vraag naar gebruik van de faciliteit dekkend is voor de capaciteit die wordt vrijgemaakt.

Bij B3 komt er ten opzichte van B2 meer capaciteit vrij voor het gebruik van AI. Aangezien de toegankelijkheid van een AI-fabriek in een andere Europese lidstaat kleiner is, is het aannemelijk dat de activiteiten in Nederland ook minder omvangrijk zullen zijn. Het effect van B2 is sterk afhankelijk van de implementatie van de hub-functie. De hub ontwikkelt (en verspreidt) expertise om de AI-faciliteit te benaderen, waarbij het technisch niet uitmaakt of de AI-faciliteit in Nederland of Finland staat. Hoewel B3 iets beter scoort, is het verschil met B2 naar verwachting niet groot.

Met B1 vindt de doorontwikkeling van kennis en vaardigheden plaats bij de selecte groep die momenteel gebruik maakt van huidige supercomputer. Deze groep bestaat voornamelijk uit wetenschappers, en daarmee is de kans groter dat er geen breder AI-ecosysteem inclusief overheid en bedrijfsleven opgebouwd kan worden.

Tabel 9. Ranking beleidsopties op het criterium Kennis & vaardigheden: Gebruik AI-faciliteiten

	3e (Slechtste)	2e	1e (Beste)
Gebruik van AI-faciliteiten	B1		B2 B3

Ontwikkeling en toepassing AI

Kennis van ontwikkeling en toepassing AI raakt aan bovenstaande laag Figuur 1 en gaat over de kennis en vaardigheden bij de eindgebruikers van de faciliteit, de onderzoekers en (door)ontwikkelaars.

Figuur 7. Kennis van ontwikkeling en toepassing AI

Net als bij de twee voorgaande criteria geldt hier dat met B3, vanwege de beschikbare capaciteit en toegankelijkheid, naar verwachting de meeste kennis en vaardigheden t.a.v. de ontwikkeling en toepassing van AI worden opgebouwd. Ook hier is dit wel afhankelijk van de mate waarin de vraag aansluit bij de capaciteit die beschikbaar komt.

Ook de investeringen in B2 hebben een versterkend effect op de kennis en vaardigheden. De wijze waarop de Nederlandse hub wordt vormgeven is hier de doorslaggevende factor; een expertisecentrum van waaruit gebruikers van de faciliteit worden ondersteund bij de ontwikkeling en toepassing van nieuwe modellen, legt meer gewicht in de schaal dan een hub waarbij partijen elkaar enkel kunnen ontmoeten en samenwerken.

Bij B1 volstaat de huidige beschikbare capaciteit niet voor de ontwikkeling van zeer complexe en geavanceerde modellen. De kennis en vaardigheden die worden opgebouwd zitten dus op het niveau van laagdrempelige modellen. Gezien de wetenschap nu hoofdzakelijk gebruik maakt van de nationale supercomputer, is het waarschijnlijk dat de kennis en vaardigheden die hierbij worden opgedaan ook vooral binnen deze groep zal neerslaan. De overheid en het bedrijfsleven zullen hier in mindere mate van profiteren.

Tabel 10. Ranking beleidsopties op het criterium Kennis & ervaring: Ontwikkeling en toepassing AI

	3e (Slechtste)	2e	1e (Beste)
Ontwikkeling en toepassing AI	B1	B2	B3

4.5.2 AI-modellen

Met een AI-fabriek kunnen verschillende AI-taken worden uitgevoerd. We onderscheiden drie verschillende taken (in navolging van [2], zie eerder paragraaf 2.2), die we hieronder uitwerken.

Nieuwe pre-trained AI-modellen

Zoals toegelicht bij het criterium *Ingezette rekenkracht*, zijn B2 en B3 voor het pre-trainen van *foundation* modellen, aangezien de huidige beschikbare rekenkracht hiervoor niet volstaat. De voorgestelde schaalvergroting met B2 en B3 maakt een substantiële sprong in de maximale omvang van de modellen mogelijk.

De publieke sector casus toont een beperkte vraag naar het ontwikkelen van *foundation* modellen bij overheidsorganisaties. Op basis van de huidige aantrekkingskracht van AI is het aannemelijk dat de (Nederlandse) wetenschap zich in toenemende mate met de ontwikkeling van deze modellen zal bezighouden. Op basis van de aangeleverde wetenschapscasus kan niet vastgesteld worden hoe groot de capaciteit hiervoor dient te zijn. In de bedrijfscasus wordt vooral de vraag voor meer rekenkracht vanuit start-ups, scale-ups en het mkb genoemd. Uit een eerdere studie naar AI-startups blijkt dat hiervan ongeveer 3% werkt aan *foundation* modellen [22].

Op basis van de vergelijkbare capaciteit die beschikbaar wordt gemaakt met B2 of B3, zijn de uitkomsten t.a.v. nieuwe pre-trained AI-modellen vergelijkbaar. Door een (waarschijnlijk) hogere toegankelijkheid (zie boven), heeft B3 wel een voorsprong. Het is dan wel noodzakelijk dat de faciliteit daadwerkelijk wordt ingezet voor de ontwikkeling van complexe en geavanceerde modellen. Dit is onzeker op basis van de aangeleverde cases en de toekomstige vraag naar *foundation* modellen. Wel is zeker dat B1 in ieder geval geen of nauwelijks zicht geeft op de ontwikkeling van deze modellen.

Tabel 11. Ranking beleidsopties op het criterium Nieuwe pre-trained AI-modellen

	3e (Slechtste)	2e	1e (Beste)	
Nieuwe pre-trained AI-modellen	B1		B2	B3

Nieuwe fine-tuned AI-modellen

Voor de *fine-tuning* van bestaande modellen is minder rekenkracht nodig en voor veel gevallen volstaat hiervoor de bestaande supercomputer, of kan gebruik worden gemaakt van commerciële cloudinfrastructuur (al kan dit om andere redenen minder aantrekkelijk zijn).

Vanuit het perspectief van rekenkracht zijn B2 en B3 nagenoeg vergelijkbaar. Ten opzichte van B1 bieden deze opties met name meer mogelijkheid om projecten parallel aan elkaar uit te voeren. Aangezien de behoefte voor *fine-tuning* duidelijk naar voren komt in de cases is het aannemelijk dat hier in de praktijk ook vraag naar is. Gelet op de benodigde rekenkracht voor dit type AI-ontwikkeling is wel de vraag of B3 hiervoor een doelmatige oplossing biedt. Met B1 komt er geen capaciteit beschikbaar voor het

finetunen van AI-modellen. Dit beperkt voornamelijk het bedrijfsleven en overheden die geen toegang hebben tot de nationale supercomputer, en daardoor zijn aangewezen op commerciële cloudinfrastructuur.

Tabel 12. Ranking beleidsopties op het criterium Nieuwe fine-tuned AI-modellen

	3e (Slechtste)	2e	1e (Beste)
Nieuwe finetuned AI-modellen	B1	B2	B3

Toepassing ontwikkelde AI-modellen

De toepassing van ontwikkelde AI-modellen, ofwel inferentie, vraagt aanzienlijk minder rekenkracht dan de ontwikkeling van deze modellen. Hiervoor biedt een AI-fabriek dus beperkt toegevoegde waarde anders dan dat er een veelvoud van toepassingen gelijktijdig op de faciliteit kunnen worden uitgevoerd. De EuroHPC JU [11] biedt echter de mogelijkheid dat de AI-fabriek wordt ingezet voor commerciële doeleinden (zie 3.9 Commercial Access in [11]). Hiervoor gelden geen restricties die voor andere doelgroepen wel gelden, zoals het feit dat voor wetenschappelijke inzet van de AI-fabriek aangetoond moet worden dat er extreme rekencapaciteit benodigd is voor een specifieke use case. Hiermee is het mogelijk dat de AI-fabriek wordt ingezet voor de toepassing van AI, ook al is dit niet de meest doelmatige inzet; een regulier datacenter of lokale faciliteiten met een beperkt aantal GPU's kunnen hiervoor ook worden ingezet. Gegeven de (waarschijnlijk) betere toegankelijkheid van B3, geldt ook hier dat B3 een (lichte) voorkeur krijgt boven B2.

Tabel 13. Ranking beleidsopties op het criterium Toepassing ontwikkelde AI-modellen

	3e (Slechtste)	2e	1e (Beste)
Toepassing ontwikkelde AI-modellen	B1	B2	B3

4.6 Outcome

4.6.1 AI-ecosysteem

De verschillende beleidsopties zullen een ander effect hebben op de ontwikkeling en versterking van het AI-ecosysteem in Nederland. De drie beleidsopties verschillen in

termen van verwachte schaalvoordelen, scopevoordelen en agglomeratievoordelen voor het AI-ecosysteem (zie eerder paragraaf 2.3).

De **schaalvoordelen** worden met name behaald bij B2 en B3. Er zijn (vrijwel) geen partijen in het bedrijfsleven, de overheid of de wetenschap die op eigen kracht een grote AI-faciliteit kunnen realiseren. Tegelijkertijd is er bij veel partijen wel behoefte om een deel van een dergelijke AI-faciliteit te gebruiken. Een 'shared facility' kan daarom de benodigde schaal creëren waardoor de schaalvoordelen ook behaald kunnen worden. Met B1 is dit niet aan de orde, omdat er geen eigen AI-fabriek is en enkel een selecte groep in competitie met andere lidstaten voor een (klein) deel van de beschikbare rekencapaciteit in aanmerking komt. Met B3 is het daarnaast denkbaar dat de gecreëerde schaalvoordelen niet enkel aantrekkelijk zijn voor Nederlandse partijen; het is goed mogelijk dat er interesse bij buitenlandse partijen bestaat en/of ontstaat om gebruik te maken van de rekenkracht die in Nederland geïnstalleerd is.

De **scopevoordelen** worden ook met name behaald bij B2 en B3. Deze scopevoordelen hebben primair betrekking op de mogelijkheid om gebruik van een grote AI-faciliteit én het ontwikkelen van AI-modellen te combineren. Dit geldt dus ook voor de benodigde kennis en kunde die op deze combinatie ontwikkeld en ingezet wordt. Met B1 kan men weliswaar (kleinere) AI-modellen ontwikkelen, maar kan dit niet of amper met behulp van grootschalige AI-rekenkracht gebeuren waardoor de mogelijkheden voor innovatie in AI-modellen beperkt zijn.

De **agglomeratievoordelen** worden ook met name bij B2 en B3 gecreëerd. In deze twee beleidsopties kunnen Nederlandse partijen experimenteren, werken en leren met de beschikbare AI-rekenkracht. Er worden (hoogstwaarschijnlijk) meer nieuwe AI-modellen ontwikkeld en er wordt meer kennis en kunde opgebouwd. Dit heeft op zijn beurt weer een verder versterkend effect op het AI-ecosysteem. Deze AI-modellen en opgebouwde kennis en kunde worden namelijk ook weer makkelijker met andere Nederlandse spelers gedeeld, er ontstaan naar verwachting meer nieuwe ideeën voor ontwikkeling van AI-modellen, en er ontstaat een grotere aantrekkingskracht voor talent en (economische) activiteit. Dit leidt vervolgens weer tot meer AI-modellen, kennis en kunde, etc. Daarnaast kunnen de ontwikkelde AI-modellen toegepast worden, en is het aannemelijk dat landen en sectoren met een grotere kennisbasis beter in staat zullen zijn om de kansen van (nieuwe) AI-modellen te herkennen en deze kansen ook te verzilveren. Nederlandse partijen met kennis en kunde van de (nieuwe) AI-modellen kunnen immers ook andere partijen in Nederland helpen met het implementeren van AI-toepassingen.

De kennis en ervaring die hiermee opgedaan wordt kunnen ook weer gedeeld worden (kennis-spillovers), waardoor zowel de vraag naar AI als het aanbod van AI versterkt wordt. Dus ook partijen die zelf geen AI-modellen ontwikkelen kunnen uiteindelijk (indirect) profiteren van een AI-fabriek. Hoewel de meningen verdeeld zijn, is onze inschatting dat deze agglomeratievoordelen niet zozeer afhankelijk lijken te zijn van de

fysieke locatie van de AI-fabriek zelf, maar met name van de *mensen en organisaties* die zich hiermee bezighouden. Zowel met B2 als B3 kan hier volop in geïnvesteerd worden. Wel is het zo dat er met B3 meer geïnvesteerd wordt in beschikbare rekenkracht waardoor er ook meer kennis en kunde ontwikkeld wordt (zie eerder), er meer kennis-spillovers zullen zijn, en de aantrekkingskracht van het ecosysteem wordt versterkt. Er is dus meer massa in B3 waardoor dit vliegwieleffect harder aangezet wordt, maar dit staat dus in feite los van de fysieke locatie waar de AI-fabriek staat. Wel kan het zijn dat er met B3 meer partijen met de AI-fabriek kunnen/mogen werken in verband met juridische restricties of organisatiebeleid, en dat het eventueel makkelijker kan zijn om data te delen binnen de landsgrenzen van Nederland.

Het aan te trekken talent en bedrijvigheid beperkt zich niet per definitie tot de Nederlandse grenzen. Buitenlandse spelers kunnen actief willen worden op de Nederlandse infrastructuur en kunnen mogelijk ook betrokken worden in het Nederlandse AI-ecosysteem waardoor het vliegwieleffect (nog) meer kracht bijgezet wordt. De mate waarin B3 leidt tot het aantrekken van buitenlandse spelers is voor ons niet goed vast te stellen, maar het is niet onwaarschijnlijk dat er interesse vanuit het buitenland zal ontstaan.

Tot slot gaat er uit de verschillende beleidsopties een ander *signaal* vanuit de overheid uit. Met B3 (en in substantiële mate nog steeds met B2) geeft de Nederlandse overheid expliciet aan om in te willen zetten op AI en is het denkbaar dat dit de aantrekkingskracht van het Nederlandse AI-ecosysteem vergroot en daarmee het beschreven vliegwieleffect verder kracht bijzet.

Innovatie in bestaande organisaties

We verwachten dat B2 en B3 tot de meeste AI-innovatie in bestaande organisaties zullen leiden. Deze innovatie komt dus enerzijds direct tot stand door het ontwikkelen van nieuwe AI-modellen door bestaande organisaties en anderzijds indirect doordat de ontwikkelde modellen en ontwikkelde kennis bij andere partijen in het ecosysteem neerslaan en ingezet kunnen worden om te innoveren. Op basis van de beschreven cases lijkt het aannemelijk dat met name bestaande grootbedrijven en bepaalde wetenschappelijke disciplines innoveren door zelf ook daadwerkelijk nieuwe AI-modellen te ontwikkelen. Hoewel er ook casussen kunnen zijn voor de overheid en het mkb, lijken deze schaarser te zullen zijn.

De AI-innovatie door (nieuwe) AI-modellen toe te passen heeft betrekking op een bredere groep organisaties in het ecosysteem, en hangt onder meer af van de toepasbaarheid van de nieuwe ontwikkelde AI-modellen, de mate van overdracht van AI-kennis en -kunde in Nederland, en de nieuwe activiteiten bij bestaande en nieuwe organisaties. Met B3 zal waarschijnlijk meer innovatie tot stand komen dan met B2, omdat er sprake is van een grotere impuls en een sterker vliegwieleffect. De mate waarin B2 en B3 van elkaar verschillen hangt met name af van de investeringen die gedaan

worden op het gebied van kennis en kunde en ecosysteemontwikkeling in brede zin, en niet zozeer van de fysieke locatie van de AI-fabriek.

Tabel 14. Ranking beleidsopties op het criterium AI-ecosysteem: Innovatie in bestaande organisaties

	3e (Slechtste)	2e	1e (Beste)
Innovatie in bestaande organisaties	B1	B2	B3

Ontwikkeling nieuwe (innovatieve) organisaties

Een AI-fabriek kan ook leiden tot nieuwe (innovatieve) organisaties. In eerste instantie hebben we het dan over start-ups en scale-ups, maar het is niet ondenkbaar dat nieuwe (innovatieve) organisaties in de wetenschap en/of de overheid gevormd worden.

Hoewel het niet mogelijk is om precies te voorspellen hoeveel start-ups er met of zonder AI-fabriek zullen zijn, is het aannemelijk dat een AI-fabriek tot meer start-ups leidt. Er is op dit moment al sprake van onvervulde vraag. Daarbij kan het beschikbaar hebben van de AI-fabriek ook een reden/voorwaarde zijn om een startup te starten. Daarnaast kan het versterken van het AI-ecosysteem in brede zin ook tot nieuwe kennis en ideeën leiden, die op hun beurt weer aan de vooravond staan van een nieuwe startup. Startups kunnen dus ook als een resultaat van het vliegwieleffect gezien worden.

Beschikbare capaciteit, toegankelijkheid van de faciliteit en een sterk AI-ecosysteem zijn het meest waarschijnlijk met B3 waardoor er naar verwachting ook de meeste start-ups zullen komen in dit scenario. Met B2 zal er naar verwachting ook een goed AI-ecosysteem opgebouwd kunnen worden met bijbehorende start-ups, maar zal het gezien de kleinere impuls naar verwachting ook tot minder start-ups leiden.

Tabel 15. Ranking beleidsopties op het criterium AI-ecosysteem: Ontwikkeling (nieuwe) innovatieve organisaties

	3e (Slechtste)	2e	1e (Beste)
Ontwikkeling (nieuwe) innovatieve organisaties	B1	B2	B3

Versterking overige ecosysteem-elementen

Zoals eerder benoemd verschillen de beleidsopties in de mate waarin ze diverse elementen van het AI-ecosysteem versterken. Naast de relatie die ze hebben tot verwachte innovatie (zie bovenstaande twee criteria), is een ander belangrijk onderscheid de impact die ze hebben op de input *Competenties & capaciteit*. Dit element is een belangrijke input voor het AI-ecosysteem en wordt door een AI-fabriek op verschillende manieren beïnvloed. Allereerst leidt het werken aan en met de AI-fabriek tot nieuwe kennis en kunde; deze opgedane kennis en kunde vloeit ook weer het AI-ecosysteem in. Daarnaast worden ook modellen en innovaties ontwikkeld, waarbij ook sprake kan zijn van kennis-spillovers naar andere partijen. De kennis en kunde bij die partijen stroomt ook weer het AI-ecosysteem in. Een derde mechanisme is de aantrekkende werking richting AI-talent in binnen- en buitenland. Door interessante activiteiten te ontplooien met de AI-fabriek én resulterende AI-activiteiten daarbuiten (bijv. toepassen van AI bij andere spelers binnen Nederland) is het aantrekkelijk voor talent om hier te komen en/of blijven werken. Ook deze aangetrokken en behouden kennis en kunde stroomt het ecosysteem weer in, waardoor deze weer verder versterkt wordt. Dit vliegwieleffect zien we als belangrijk element in de meerwaarde van de AI-fabriek. In lijn met de eerdere argumentatie zal met B3 de meeste versterking van de *human capital* component plaatsvinden, en zal B2 hierop volgen. Met B1 gebeurt er relatief weinig (nieuws) en zal het beschreven vliegwieleffect niet of minder op gang komen.

Een ander element van het ecosysteem dat versterkt kan worden is de input (training)data. Wanneer er meer capaciteit van een AI-fabriek gebruikt wordt en kan worden is het waarschijnlijker dat een deel van de gebruikte data ook gedeeld zou kunnen worden met andere partijen. Hierdoor worden de mogelijkheden voor modelontwikkeling verbreed en kan dit op zijn beurt weer leiden tot additionele innovatie.

Tabel 16. Ranking beleidsopties op het criterium AI-ecosysteem: Versterking overige ecosysteemelementen

	3e (Slechtste)	2e	1e (Beste)
Versterking overige ecosysteemelementen	B1	B2	B3

4.6.2 Verdere outcomes

Open strategische autonomie

Op het moment zijn Nederland en de EU met name afhankelijk van Amerikaanse partijen die grote *foundation* modellen ontwikkelen; de Big Tech zoals OpenAI en Google.

Een AI-fabriek kan helpen om de afhankelijkheid van dergelijke partijen te verkleinen, ook al is het niet aannemelijk dat we binnen de EU op korte termijn onze volledige technologische achterstand op dit gebied weg gaan werken. Desalniettemin kan het in eigen beheer ontwikkelen van belangrijke en fundamentele AI-modellen van belang zijn voor het creëren van meer (open) strategische autonomie en een strategische voorsprong op specifieke AI-toepassingen.

Op Europees niveau is het de vraag of het veel uitmaakt of de AI-fabriek in Finland of Nederland staat; vermoedelijk niet. Toegang tot een AI-fabriek onder EU-controle is het belangrijkste en dat wordt met zowel B2 als B3 geregeld. Wel is een belangrijk verschil dat er met B3 waarschijnlijk méér in AI geïnvesteerd wordt; iets dat de EU in generieke zin ten goede zou kunnen komen.

Op nationaal niveau zal B3 tot meer (open) strategische autonomie leiden, omdat een eigen faciliteit de meeste controle biedt en het (in sommige situaties) makkelijker kan zijn om compliant te zijn. Er kunnen (overheids)casussen zijn waarbij strikt geëist wordt dat de data op Nederlandse bodem blijven. Tegelijkertijd zien we in de praktijk dat men bij dergelijke 'gevoelige' casussen aangeeft dat data überhaupt niet de organisatiemuren mag verlaten. Voor de wetenschap maakt het niet veel uit waar de AI-fabriek staat.

Samengevat leiden zowel B2 en B3 tot meer (open) strategische autonomie, en is het verschil tussen B2 en B3 beperkt.

Tabel 17. Ranking beleidsopties op het criterium Open strategische autonomie

	3e (Slechtste)	2e	1e (Beste)
Open strategische autonomie	B1		B2 B3

Maatschappelijk vertrouwen in AI

De verschillende beleidsopties kunnen een verschillend effect hebben op het maatschappelijke vertrouwen in AI. Dit thema is niet expliciet teruggekomen in de cases.

Hoewel het vanuit onze positie moeilijk te onderbouwen is, is het onzes inziens aannemelijk dat het maatschappelijk vertrouwen in AI het grootst zal zijn als we (1) als land hier zelf kundig in zijn, (2) er zelf mede vorm aan kunnen geven en (3) het slim en verantwoord kunnen inzetten binnen NL. Deze facetten zullen het meest versterkt worden met B3, in mindere mate met B2 en het minst met B1. Daarbij is de kans groter dat er met B3 (en in iets mindere mate met B2) een hoogwaardiger AI-ecosysteem ontstaat wat ook bij kan dragen aan het maatschappelijk vertrouwen.

Tabel 18. Ranking beleidsopties op het criterium Open strategische autonomie

4.7 Impact

In de laatste stap van ons analysekader kijken we naar de lange termijn veranderingen die verwacht kunnen worden met de beleidsinzet. Zoals eerder gesteld is deze impact niet meetbaar of anderszins herleidbaar tot de acties, maar geeft dit niveau de context weer *waarom* doelstellingen worden nagestreefd. In het kader van de AI-fabriek betreft dit twee impactthema's: het lange termijn-verdienvermogen en de bijdrage aan oplossingen voor maatschappelijke vraagstukken.

Lange termijn verdienvermogen

De impact van een AI-fabriek, gericht op het pre-trainen van AI-modellen, op het lange termijn verdienvermogen van Nederland kan worden gedefinieerd als het vermogen om via deze modellen economische waarde te creëren door innovatie en efficiëntieverbeteringen. Door AI-modellen die specifieke problemen oplossen te ontwikkelen en toepassen, kunnen bedrijven en overheden hun productiviteit verhogen, kosten verlagen, en nieuwe producten en diensten introduceren. In een eerder rapport beschreven we al dat AI een belangrijke ontwikkeling is binnen en voor Nederlandse groeimarkten [26]. Dit stimuleert op haar beurt economische groei, opent nieuwe markten, en verhoogt het concurrentievermogen van Nederland op internationaal niveau ('voortrekkersrol'), wat essentieel is voor het waarborgen van langdurig verdienvermogen.

De bijdrage aan het lange termijn verdienvermogen verwachten we met name vanuit het wetenschappelijk en bedrijfsmatig gebruik (wetenschaps- en bedrijfscasus) en dus in minder mate vanuit de overheid (publieke sector casus).

Allereerst zou een toename aan wetenschappelijke output kunnen leiden tot economische baten. Het ontwikkelen van de modellen zit wel vroeg in de innovatieketen (lage TRL), dus het pad naar succesvolle innovatie en adoptie is nog onzeker. Zaak is dus dat de ontwikkelde kennis kan doorstromen naar bedrijven. Ook het aantrekken en behouden van studenten leidt op termijn tot meer arbeidspotentieel, wat een positieve bijdrage kan hebben op de innovatiekracht en het vestigingsklimaat voor AI-gerelateerde bedrijvigheid (toegang tot talent).

Met de AI-fabriek en het aanpalende AI-ecosysteem is het aannemelijk dat er meer nieuwe startups gecreëerd worden en/of dat die door kunnen groeien naar scale-ups. Ook worden bestaande bedrijven versterkt m.b.v. een concurrentievoordeel. B3 zal daar het meeste aan kunnen bijdragen, zoals hierboven uitgewerkt. Een vraag is wel er vanuit het ecosysteem zelf bereidheid zal ontstaan om in de faciliteit te (her)investeren (door het aanbod voor bedrijven en de andere stakeholders zo aantrekkelijk te maken dat de structurele meerwaarde duidelijk wordt), of dat de financiering (al dan niet volledig) vanuit een publieke bron moet blijven komen (wat past bij een karakter van een gedeelde onderzoeks- en ontwikkelfaciliteit).

Het is onze verwachting dat B3 tot de meeste AI-ecosysteemvorming en de meeste innovatie leidt. Ook is het onze verwachting dat AI een belangrijke driver (wellicht de belangrijkste) van economische groei gaat worden in de komende decennia, waardoor B3 dus het beste scoort op het lange termijn-verdienvermogen. Indien er met B2 substantieel wordt geïnvesteerd in de AI-hub, kan een groot gedeelte van deze voordelen ook worden behaald met B2.

Tabel 19. Ranking beleidsopties op het criterium Langetermijn verdienvermogen

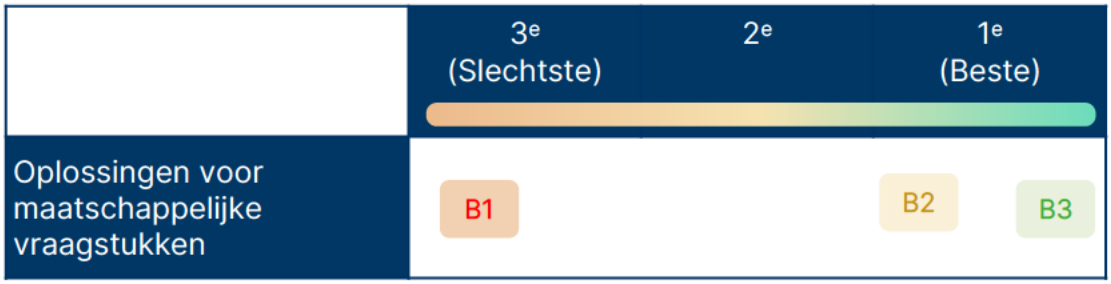
	3e (Slechtste)	2e	1e (Beste)
Langetermijn-verdienvermogen	B1	B2	B3

Oplossingen voor maatschappelijke vraagstukken

De resultaten van op AI-gebaseerd onderzoek kunnen grote effecten hebben in de aanpak van maatschappelijke uitdagingen als de houdbaarheid van de zorg, energietransitie, omgaan met impact klimaatverandering, emissievrije mobiliteit, overheidsdienstverlening, circulaire landbouw en het onderwijs van de toekomst. Dergelijke maatschappelijke vraagstukken hebben een directe invloed op het welzijn van burgers.

In het verlengde van het lange termijn-verdienvermogen verwachten we dat de grootste inzet (oftewel B3) tot de meeste maatschappelijke oplossingen zou kunnen leiden. Dit geldt zowel voor het weterschappelijke onderzoek als voor de innovatieve bedrijvigheid. Indirect kan hierbij de dienstverlening van de overheid beter tegemoet komen aan maatschappelijke behoeftes. Indien er met B2 substantieel wordt geïnvesteerd in de AI-hub, kan een groot gedeelte van deze voordelen ook worden behaald met B2.

Tabel 20. Ranking beleidsopties op het criterium Oplossingen voor maatschappelijke vraagstukken



4.8 Randvoorwaarden

In deze paragraaf kijken we naar de aansluiting van de drie beleidsopties op de noodzakelijke randvoorwaarden voor het opzetten van een AI-fabriek. Als niet aan deze randvoorwaarden wordt voldaan, kan de verandering niet op de gewenste manier plaatsvinden. Overkoepelend kan gesteld worden dat geen van de drie beleidsopties uitgesloten dient te worden op basis van de randvoorwaarden, maar dat de opties wel van elkaar verschillen in de mate waarop ze bij de randvoorwaarden aansluiten.

Wet- en regelgeving

In de aangeleverde cases wordt wet- en regelgeving vrijwel uitsluitend benoemd bij het delen en beschikbaar stellen van data. Het gros van de partijen geeft aan zich hierbij te richten op Europese wet- en regelgeving op het vlak van data- en cybersecurity. Vanuit die optiek is er strikt genomen geen verschil tussen een faciliteit in een andere Europese lidstaat (B2) of een nationale AI-fabriek (B3), omdat beiden worden gedekt door Europese wet- en regelgeving. Er zijn echter ook partijen die aangeven dat een faciliteit in Nederland meer grip geeft op de ontwikkeling van AI in overeenstemming met de eisen van Nederlandse organisaties (en bijbehorende juridische kaders). B2 kan dus mogelijk enkele uitdagingen geven ten aanzien van wet- en regelgeving. Met B1 worden geen nieuwe activiteiten ontplooid, wat dus ook geen vraagstukken ten aanzien van wet- en regelgeving meebrengt.

Tabel 21. Ranking beleidsopties op de randvoorwaarde Wet- en regelgeving

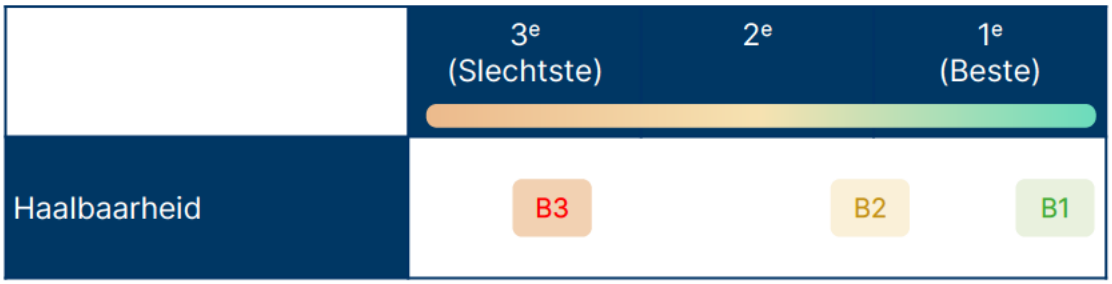
Haalbaarheid

Op het vlak van haalbaarheid zitten de grootste knelpunten duidelijk bij B3. SURF loopt zonder nadere afspraken grote risico's bij B3 (financieel – onderschatting van kosten

en prijsontwikkeling, juridisch – BTW-afspraken en inbestedingsmodel en capaciteit – beslag op resources en management) [20]. Daarnaast zijn er ook risico's t.a.v. de stroomvoorziening en -prijs bij deze optie. Deze risico's kunnen weliswaar gemitigeerd worden met garanties van overheid, leden of verzekeraars; zonder deze mitigerende maatregelen zijn dit grote knelpunten voor de haalbaarheid van deze beleidsoptie.

Ook voor B2 moet een stevige bijdrage geleverd worden, zowel in geld als qua menskracht. De omvang van deze benodigde investering is daarentegen significant minder risicovol dan B3 en daarmee is de haalbaarheid van deze optie ook beter [20]. B1 heeft vanuit het perspectief van risicomanagement de beste papieren en is daarmee zeer haalbaar, maar dat het belangrijkste risico van deze optie de dreiging is dat Nederland de boot mist voor wat betreft de ontwikkeling van AI [20].

Tabel 22. Ranking beleidsopties op de randvoorwaarde Haalbaarheid



Commitment van betrokkenen

De beleidsopties verschillen in de mate waarin wetenschappers, het bedrijfsleven en de overheid betrokken kunnen worden bij ontwikkeling van AI, en *committed* moeten zijn voor de uitwerking van de beleidsoptie.

Vanuit het perspectief van **ecosysteemvorming** is het wenselijk dat het netwerk dat zich hiermee bezighoudt sterk wordt uitgebreid om optimale kruisbestuiving te hebben tussen wetenschap, bedrijfsleven en overheden. De opzet van een nationale AI-fabriek bij B3 (voor de weging van commitment van Europese partners, zie *Europese samenwerking*) biedt hiervoor de meeste kansen, maar gelet op de huidige vraag vanuit deze partijen naar de ontwikkeling van geavanceerde modellen is het twijfelachtig of de commitment in de praktijk daadwerkelijk zo groot is. Op basis van de aangeleverde cases is daar geen eenduidig antwoord op te geven, maar duidelijk is dat er een risico tot onderbenutting van de beschikbare capaciteit met B3 is.

Bij B1 is er enkel commitment nodig van de partijen die reeds gecommitteerd zijn aan de bestaande middelen en faciliteiten. Er is dus geen risico ten aanzien van commitment, maar er wordt ook geen extra commitment aangegaan binnen het AI-ecosysteem.

Met B2 komt er extra capaciteit beschikbaar op het vlak van beschikbare rekenkracht. Daarnaast hebben de betrokken extra expertise tot hun beschikking vanuit de

Nederlandse hub. Daarmee wordt de betrokkenen de mogelijkheid geboden om zich sterker te verbinden met de ontwikkeling van AI, zonder dat daarbij het risico van onderbenutting wordt gelopen in de omvang waarin dat speelt bij B3.

Commitment speelt ook voor de **governance** van een AI-fabriek. Met B1 is er geen governance benodigd, maar zijn partijen afhankelijk van de keuzes en ontwikkelingen in andere Europese landen. Met B2 is een beperkte governance benodigd om de rekentijd voor het Nederlandse deel te beheren. Er is echter beperkte invloed op het ontwerp en gebruik van de gehele AI-fabriek [20], doordat de governance over de AI-fabriek in een andere Europese lidstaat is georganiseerd. Met B3 is een stevige governance benodigd voor het beheer van de AI-fabriek, rekentijd en financiën. Een dergelijke governance kent ten minste twee uitdagingen. Er moet voldoende steun zijn onder de leden van SURF (als penvoerder van een Nederlandse AI-fabriek) voor de gehele looptijd, terwijl niet alle leden in dezelfde mate gebruik zullen maken van de AI-fabriek [20]. Daarnaast dient de governance in een publiek-private samenwerking te worden vormgegeven, een model dat nog niet goed verkend is.

Tabel 23. Positionering van de beleidsopties t.o.v. Commitment van betrokkenen

	3e (Slechtste)	2e	1e (Beste)
Commitment van betrokken	B3	B1	B2

Behouden positie nationale supercomputer

De laatste randvoorwaarde bij de drie beleidsopties is het behoud van de positie van de bestaande nationale supercomputer, Snellius (zie het tekstkader op de volgende pagina) en toekomstige opvolgers. In onze optiek gaat B3 gepaard met het grootste risico voor Snellius, aangezien de beschikbare rekenkracht in de nationale AI-fabriek de bestaande supercomputer sterk voorbijstreeft. Dit zal ongetwijfeld een deel van de huidige gebruikers van Snellius naar de AI-fabriek trekken. Aangezien Snellius niet geschikt is voor de ontwikkeling van *foundation* modellen, is er op dat vlak geen reëel risico voor overlap. Wat betreft beheer is er het risico dat SURF onvoldoende capaciteit heeft om zowel een AI-fabriek als de nationale supercomputer te beheren [20], en er vanuit de overheid onvoldoende middelen zijn om twee supercomputers financieel te onderhouden [16].

Met B1 blijft Snellius behouden, maar zal de huidige supercomputer niet lang kunnen voorzien in de toenemende vraag naar rekenkracht. Op het vlak van AI-ontwikkeling zullen potentiële gebruikers eerder uitwijken naar alternatieven, zoals LUMI, waarmee de positie van de nationale supercomputer op lange termijn minder sterk wordt.

De hub die met B2 wordt opgezet ter ondersteuning van AI-ontwikkeling en het beschikbaar stellen van (schaarse) expertise biedt in potentie *spillover*-effecten voor Snellius. Het is denkbaar dat er situaties ontstaan waarin AI-ontwikkelaars gebruik zullen maken van de expertise in Nederland, maar vanwege beschikbaarheid van data de ontwikkeling niet in een andere lidstaat (zoals Finland) willen uitvoeren. Voor projecten die qua schaal op Snellius kunnen worden uitgevoerd zou dit mogelijkheden kunnen scheppen.

Tabel 24. Ranking beleidsopties op de randvoorwaarde Behouden positie nationale supercomputer

	3e (Slechtste)	2e	1e (Beste)
Behouden positie nationale supercomputer	B3	B1	B2

Wat er nu is: de Snellius-supercomputer

Snellius is een in 2021 in gebruik genomen supercomputer ten behoeve van Nederlandse universiteiten en onderzoeksinstellingen. Snellius is aangeschaft en wordt beheerd door SURF en bevindt zich fysiek in Amsterdam. De supercomputer had op moment van ingebruikname een rekensnelheid van initieel 6 petaflop/s[6] en uiteindelijk 29 petaflop/s, en een opgenomen piekvermogen van 1,4 megawatt. De realisatie van de supercomputer kostte € 20 miljoen. De beoogde levensduur van Snellius was 5 jaar [27].

De Snellius-supercomputer bestaat uit een groot aantal (initieel ongeveer 77.000, uiteindelijk ongeveer 220.000) traditionele rekenkernen (CPU's). Daarnaast beschikt het systeem sinds kort over 408 GPU's [28]. Het trainen van een grootschalig taalmodel zoals GPT-3 van OpenAI (175 miljard parameters) kost naar schatting vier maanden.

[6] De eenheid 'flop' verwijst naar 'floating point operation', een wiskundige bewerking op zwevendekommagetallen van 64 bit. Een petaflop staat voor 10^{15} van dergelijke operaties. Het aantal petaflop/s geeft het aantal operaties per seconde en is daarmee een indicatie van de snelheid van de computer. Een petaflop-dag (een dag lang rekenen met 1 petaflop/s) komt overeen met $8{,}64 * 10^{19}$ uitgevoerde wiskundige operaties.

4.9 Overzicht resultaten analysekader

Tabel 25 toont een overzicht van de resultaten van de meta-analyse, waarbij de verschillende beleidsopties per onderdeel van de *theory of change* (Figuur 4) zijn beoordeeld en gerangschikt.

Tabel 25. Overzicht resultaten meta-analyse: ranking van beleidsopties per onderdeel van de theory of change en randvoorwaarden

Onderdelen van de Theory of Change	3e	2e	1e
Input			
Bijdrage EU	B1	B2	B3
Bijdrage NL	B3	B2	B1
Huisvesting & energie	B3		B1 B2
Competenties & capaciteit	B1		B2 B3
(Trainings)data	B1	B2	B3
Output			
Gebouwde AI-fabriek	(Niet gescoord)		
Toegankelijkheid AI-faciliteiten	B1	B2	B3
Europese samenwerking	B1	B3	B2
Ingezette rekenkracht	B1		B2 B3
Int. outcome			
Kennis bouw, onderhoud, beheer AI-faciliteiten	B1	B2	B3
Kennis van gebruik AI-faciliteiten	B1	B2	B3

Onderdelen van de Theory of Change	3e	2e	1e
Int. outcome			
Kennis van ontwikkeling en toepassing AI	B1	B2	B3
Nieuwe pre-trained AI-modellen	B1	B2	B3
Nieuwe fine-tuned AI-modellen	B1	B2	B3
Toepassing ontwikkelde AI-modellen	B1	B2	B3
Outcome			
Innovatie in bestaande organisaties	B1	B2	B3
Ontwikkeling nieuwe (innovatieve) organisaties	B1	B2	B3
Versterking overige ecosysteem-elementen	B1	B2	B3
Strategische autonomie	B1	B2	B3
Maatschappelijk vertrouwen in AI	B1	B2	B3
Impact			
Langetermijn verdienvermogen	B1	B2	B3
Oplossingen maatschappelijke vraagstukken	B1	B2	B3

Randvoorwaarden	3e	2e	1e
Wet- en regelgeving		B2	B1 B3
Haalbaarheid	B3	B2	B1
Commitment van betrokkenen	B3	B1	B2
Behouden positie nationale supercomputer	B3	B1	B2

Legenda

B1	**Beleidsoptie 1.**	Bestaande middelen en instrumenten gebruiken
B2	**Beleidsoptie 2.**	Meer investeren in deelname in toekomstige Europese AI-fabrieken met een hub in Nederland
B3	**Beleidsoptie 3.**	Een AI-fabriek realiseren in Nederland

dialogic

5 Conclusies

In dit hoofdstuk geven we onze conclusies per beleidsoptie, en benoemen we een aantal aandachtspunten voor het vervolg.

5.1 Conclusies per beleidsoptie

In deze paragraaf werken we per beleidsoptie de belangrijkste voor- en nadelen uit en vatten we per beleidsoptie de consequenties samen voor (1) rekenkracht, (2) data en (3) expertise.

Beleidsoptie B1. Bestaande middelen en instrumenten gebruiken

De belangrijkste **voordelen** van de bestaande middelen en instrumenten gebruiken is dat er een beperkte behoefte is aan (financiële) bijdragen, en het weinig risico's meebrengt.

De belangrijkste **nadelen** zijn dat er geen investeringen worden gedaan in rekenkracht, ondanks behoeften vanuit met name wetenschap en bedrijfsleven. Ook wordt er geen additionele expertise en capaciteit opgebouwd ten aanzien van het beheer en gebruik van AI-faciliteiten. Toekomstige *foundation* modellen kunnen waarschijnlijk niet of zeer beperkt binnen Nederland worden ontwikkeld, waardoor het Nederlandse AI-ecosysteem afhankelijk wordt van ontwikkelingen buiten Nederland. Binnen de Europese context waar meerdere landen inzetten op AI-fabrieken via EuroHPC JU, geeft de Nederlandse overheid een signaal af dat zij geen rol voor zichzelf ziet in het Europese AI-ecosysteem. Op lange termijn brengt dit risico's met zich meer voor het Nederlandse AI-ecosysteem en het verdienvermogen van de Nederlandse economie.

Tabel 26 vat de impact op de verschillende kenmerken van een AI-fabriek met B1 samen.

Tabel 26. Impact van B1 op de kenmerken van een AI-fabriek

Kenmerk	Score	Uitleg
Rekenkracht	Laag	Er komt geen extra rekenkracht beschikbaar binnen Nederland. Europees kan in concurrentie rekenkracht worden aangevraagd bij andere AI-fabrieken.
Data	Laag	Er komt geen datacentrum beschikbaar waar grootschalige datasets kunnen worden opgeslagen, en verwerkt voor/door AI-modellen. Data dient bij commerciële aanbieders te worden verzameld en verwerkt, of bij AI-fabrieken elders in Europa.

Kenmerk	Score	Uitleg
 Expertise	Laag	Er wordt geen extra expertise opgebouwd ten aanzien van AI-faciliteiten of ontwikkeling en toepassing van AI-modellen.

Beleidsoptie B2. Meer investeren in deelname in toekomstige Europese AI-fabrieken met een hub in Nederland

De belangrijkste **voordelen** van investeren in deelname in toekomstige Europese AI-fabrieken met een hub in Nederland zijn de continuering en intensivering van Europese samenwerking (bijvoorbeeld binnen het LUMI-consortium), vergroting van de beschikbare rekenkracht voor Nederlandse gebruikers, de mogelijkheid voor Nederlandse gebruikers om *foundation* modellen te ontwikkelen, en dat zonder belasting van het Nederlandse energienet.

De belangrijkste **nadelen** zijn de grotere uitdagingen op dataverwerking binnen de wettelijke kaders of afspraken met rechthebbenden (data moet verwerkt kunnen worden buiten de landgrenzen) en minder toegankelijkheid op de rekenkracht ten opzichte van beleidsoptie B3. B2 heeft mogelijk een minder sterk vliegwieleffect voor het Nederlandse AI-ecosysteem.

Deze voordelen van B2 worden in belangrijke mate bepaald door de investering in een hub. Het is ook mogelijk om met B2 enkel aan te sluiten bij een Europees consortium, zonder Nederlandse hub. In dat geval wordt er minder expertise opgebouwd binnen Nederland, is de Europese samenwerking minder intensief (door beperkte activiteiten binnen Nederland), en is de toegankelijkheid van de Europese AI-fabriek lager door gebrek aan ondersteuning binnen Nederland.

Tabel 27 vat de impact op de verschillende kenmerken van een AI-fabriek met B2 samen.

Tabel 27. Impact van B2 op de kenmerken van een AI-fabriek

Kenmerk	Score	Uitleg
 Rekenkracht	Medium / Hoog	Er komt extra rekenkracht beschikbaar voor Nederland. Er komt mogelijk meer rekenkracht beschikbaar dan met B3 door schaalvoordelen van samenwerking in een internationaal consortium.
 Data	Medium	Er komt een datacentrum beschikbaar waar grootschalige datasets kunnen worden verwerkt voor/door AI-modellen. Daardoor is geen afhankelijkheid van commerciële aanbieders voor AI-verwerking van data. Bepaalde use cases worden niet

Kenmerk	Score	Uitleg
		ondersteund doordat data niet buiten de landgrenzen mag worden verwerkt.
 Expertise	Hoog	Er wordt een hub opgebouwd met extra expertise ten aanzien van ontwikkeling en toepassing van AI-modellen. Dit is de hoogwaardige expertise die het meest van belang is voor het AI-ecosysteem. In het geval dat alleen wordt aangesloten op een internationaal consortium zonder investeringen in een nationale hub kan de score op **Laag** uitkomen.

Beleidsoptie B3. Een AI-fabriek realiseren in Nederland

De belangrijkste **voordelen** van het realiseren van een AI-fabriek in Nederland is dat deze de beste mogelijkheden biedt voor (training)data, additionele rekenkracht biedt, de hoogste toegankelijkheid geeft en leidt tot ontwikkeling van expertise op het gebied van beheer en gebruik van AI-faciliteiten. *Foundation* modellen kunnen *binnen* Nederland geprioriteerd en ontwikkeld worden. De overheid geeft met een keuze voor beleidsoptie B3 het sterkste signaal af dat zij het Nederlandse AI-ecosysteem wil stimuleren en een stevige positie geven in het Europese AI-ecosysteem.

De belangrijkste **nadelen** zijn dat dit verreweg de duurste beleidsoptie is, met de grootste risico's op financiële dekking en governance. Een eigen AI-fabriek geeft de grootste belasting van het Nederlandse energienet. Er is een risico op verdringing van de nationale supercomputer (financieel, in beheer en in gebruik). Voor de bedrijfscasus is er een risico dat behoefte en commitment ondanks de investeringen toch tegenvallen.

Deze voor- en nadelen worden in belangrijke mate bepaald door de grootte van de investering, en daarmee de grootte van de AI-fabriek. Het is binnen de voorwaarden van de EuroHPC JU-call mogelijk om middelen aan te vragen voor **een 'kleine' AI-fabriek**. De genoemde risico's zijn dan aanzienlijk lager. Hiermee komt echter ook minder rekenkracht beschikbaar en zal het vliegwieleffect voor het Nederlandse AI-ecosysteem kleiner zijn (waarbij B2 bij gelijke middelen mogelijk leidt tot meer beschikbare rekenkracht en een sterker vliegwieleffect). Zoals hierboven benoemd, verwachten we geen additioneel vliegwieleffect door de fysieke locatie van de AI-fabriek. Het belangrijkste voordeel van een 'kleine' AI-fabriek ten opzichte van B2 is de hogere mate van toegankelijkheid. Daarbij zullen mogelijk een aantal use cases gefaciliteerd worden van partijen waarvan data niet buiten de landgrenzen mag worden verwerkt.

Tabel 28 vat de impact op de verschillende kenmerken van een AI-fabriek met B2 samen.

Tabel 28. Impact van B3 op de kenmerken van een AI-fabriek

Kenmerk	Score	Uitleg
 Rekenkracht	Hoog	Er komt extra rekenkracht beschikbaar voor Nederland. Een Nederlandse partij beheert de rekenkracht en prioriteert aanvragen. In het geval van een 'kleine' AI-fabriek kan de score op Medium uitkomen, door minder beschikbare rekenkracht.
 Data	Hoog	Er komt een datacentrum beschikbaar waar grootschalige datasets kunnen worden verwerkt voor/door AI-modellen. Daardoor is er geen strategische afhankelijkheid van commerciële aanbieders of buitenlandse AI-fabrieken voor AI-verwerking van data. Waarschijnlijk worden niet alle use cases ondersteund doordat data niet buiten de muren van instellingen mag worden verwerkt, maar een AI-fabriek binnen Nederland maakt het mogelijk hierover het gesprek aan te gaan.
 Expertise	Hoog	Er wordt een hub opgebouwd met extra expertise ten aanzien van ontwikkeling en toepassing van AI-modellen. Ook wordt er expertise opgebouwd ten aanzien van het bouwen en beheren van AI-infrastructuur.

5.2 Aandachtspunten voor het vervolg

Voor de meta-analyse zijn we uitgegaan van de drie beleidsopties zoals eerder gepresenteerd in de Kamerbrief (zie paragraaf 1.2) [1]. Hoewel dit de aanpak van de meta-analyse helpt te structureren (er zijn duidelijke verschillen tussen de beleidsopties), zijn de uiteindelijke effecten afhankelijk van de implementatie. Ten aanzien van het vervolg geven we de volgende aandachtspunten mee die naar boven zijn gekomen in deze meta-analyse:

- Beleidsoptie B2 heeft de grootste waarde bij een serieuze investering in de hub en in het internationale consortium. B2 kan in theorie geïmplementeerd worden met een minimale investering, maar veel van de voordelen en opbrengsten die we in kaart brengen in dit rapport vervallen dan. Gesteld kan worden dat bij een minimale investering in B2, deze nagenoeg samenvalt met B1 (zie ook [20]). Met een grote investering in B2 valt deze meer samen met B3 (meer expertise en meer rekenkracht, zonder de risico's van B3).
- De toegankelijkheid (de hoeveelheid rekentijd en de snelheid waarmee een aanvraag kan worden toegekend en ingepland) van een AI-fabriek met beleidsoptie B2 is sterk afhankelijk van de afspraken die worden gemaakt in het internationaal consortium. De mate waarin B2 minder goed scoort op toegankelijkheid dan B3 is dus

enigszins onzeker en afhankelijk van de afspraken die gemaakt moeten worden binnen B2.

- Beleidsoptie B3 biedt de grootste mogelijke opbrengsten, maar brengt ook de grootste risico's met zich mee. Het is de vraag of de meerwaarde van B3 ten opzichte van B2 en de benodigde middelen voor B3 daarmee doelmatig is en blijft. Daarentegen is het niet altijd mogelijk om R&D op volledig doelmatige wijze te stimuleren [29].[7]
- Als er middelen worden gereserveerd voor de implementatie van beleidsopties B2 of B3 worden idealiter niet alleen incidentele middelen vrij gemaakt om te voldoen aan de EuroHPC JU-call, maar wordt een meerjarenbegroting opgesteld voor afschrijving, beheer en vervanging van AI-faciliteiten. Binnen een dergelijke meerjarenbegroting kan ook aandacht worden besteed aan de (opvolging van de) nationale supercomputer en andere digitale infrastructuur.
- Om methodologische redenen zijn de drie beleidsopties afzonderlijk geanalyseerd. Een combinatie van (onderdelen van) beleidsopties biedt mogelijk kansen om sterktes te combineren en risico's of zwaktes te mitigeren.
- De meerwaarde van een AI-fabriek voor het AI-ecosysteem wordt ten dele bepaald door imagovorming van Nederland als hightech land. Beleidsopties B2 (met een hub) en B3 (met een eigen faciliteit) bieden kansen.

[7] Zo schrijft de high-level expert groep in haar recente rapport over het Europese kaderprogramma dat de toekomstige kansen van innovatiebeleid beperkt blijven zolang Europese en nationale overheden liever te weinig dan te veel investeren in R&D [38].

Verwijzingen

[1] Minister van Economische Zaken en Klimaat (2024). ***Kamerbrief verkenning mogelijkheden AI-faciliteit*** [www.rijksoverheid.nl] Den Haag: Ministerie van Economische Zaken en Klimaat.

[2] Rijksoverheid (2024). ***Overheidsbrede visie Generatieve AI*** [www.rijksoverheid.nl] Den Haag: Ministerie van Binnenlandse Zaken en Koninkrijksrelaties.

[3] EuroHPC JU (2024). ***The 'AI Factories' Amendment to the EuroHPC JU Regulation Enters Into Force*** [eurohpc-ju.europa.eu]

[4] Europese Raad (2024). ***Verordening (EU) 2024/1732 tot wijziging van Verordening (EU) 2021/1173 wat betreft een EuroHPC-initiatief voor start-ups om Europees leiderschap op het gebied van betrouwbare artificiële intelligentie te stimuleren*** [eur-lex.europa.eu] Brussel,

[5] John Peddle Research (2024). ***Shipments of graphics add-in boards decline in Q1 of 24 as the market experiences a return to seasonality*** [www.jonpeddie.com] Tiburon, CA,

[6] Kharpal, A. (2024). ***China seeks a homegrown alternative to Nvidia — these are some of the companies to watch*** [www.cnbc.com]

[7] van Miltenburg, O. (2021). ***Google claimt dat tpu v4-pods exaflops aan rekenkracht bieden*** [tweakers.net]

[8] Apple (2022). ***Deploying Transformers on the Apple Neural Engine*** [machinelearning.apple.com]

[9] Dialogic & PB7 (2022). ***De waarde van datacenters in regionaal perspectief***

[10] Nature (2024). ***Chemistry Nobel goes to developers of AlphaFold AI that predicts protein structures*** [www.nature.com]

[11] EuroHPC JU (2024). ***Access Policy of the EuroHPC Joint Undertaking, v2.1*** [eurohpc-ju.europa.eu]

[12] Dialogic, (2020). ***Onderzoeks- en innovatie ecosystemen in Nederland. Achtergrondstudie bij de kabinetsstrategie 'Versterken van onderzoeks- en innovatie-ecosystemen'*** [www.rijksoverheid.nl] Den Haag: Ministerie van Economische Zaken en Klimaat en het Ministerie van Onderwijs, Cultuur en Wetenschap.

[13] NL AIC (2024). ***About NL AIC*** [nlaic.com]

[14] Dialogic (2023). ***Het belang van digitale infrastructuur voor de Nederlandse digitale knooppuntrol*** [dialogic.nl] Utrecht: Dialogic.

[15] European Commission's Group of Chief Scientific Advisors (2024). ***Successful and timely uptake of artificial intelligence in science in the EU: scientific opinion***

[16] NWO & SURF (2024). ***Computational Needs for Accelerated Scientific Discovery***

[17] Algemene Rekenkamer (2024). ***Focus op AI bij de rijksoverheid*** [www.rekenkamer.nl]

[18] KplusV (2024). ***Impact-analyse AI Factory***Groningen: NOM/Provincie Groningen.

[19] RVO (2023). ***Jaarverslag WBSO 2023 - Focus Op research & development*** [www.rijksoverheid.nl] Den Haag: Ministerie van Economische Zaken en Klimaat.

[20] Stratix (2024). ***Risicoanalyse AI-factory SURF. Vergelijking van 3 scenario's***Hilversum: SURF BV.

[21] SURF (2024). ***Memo. Input scenario's AI-factory***

[22] appliedAI (2024). ***Generative AI in the European Startup Landscape 2024***

[23] SURF. ***Toegang tot rekendiensten*** [www.surf.nl]

[24] Europese Commissie (2024). ***The future of European competitiveness – A competitiveness strategy for Europe*** [commission.europa.eu] Brussel: Europese Commissie.

[25] LUMI (2024). ***LUMI consortium*** [lumi-supercomputer.eu]

[26] Dialogic & SEO (2023). ***Groeimarkten voor Nederland*** [www.rijksoverheid.nl]

[27] CWI (2021). ***Hoera! Een nieuwe nationale supercomputer: Snellius*** [www.cwi.nl]

[28] Hofmans, T. (2024). ***Snellius-supercomputer krijgt met Nvidia H100-uitbreiding 38 petaflops snelheid*** [tweakers.net]

[29] Heide, M.d. (2024). ***Behalen R&D-doelen vereist extra publieke middelen*** [esb.nu] ESB.nu.

[30] Ministerie van Buitenlandse Zaken (2024). ***Fiche 3: Verordening supercomputerinitiatief kunstmatige intelligentie*** [www.rijksoverheid.nl]

[31] Ministerie van Buitenlandse Zaken (2024). ***Fiche 4: Mededeling stimuleren van startups en innovatie in betrouwbare AI*** [www.rijksoverheid.nl]

[32] van Miltenburg, O. (2021). ***Snellius, de nationale supercomputer. 76.832 AMD-cores en 144 Nvidia A100-gpu's*** [tweakers.net]

[33] Brown, T.R. M. B., Ryder, N., en al., e. (2020). ***Language Models are Few-Shot Learners*** [arxiv.org]

[34] TOP500.org (2024). ***Top 500 June 2024*** [top500.org]

[35] NEBIUS (2023). ***Vacancy DWDM engineer*** [careers.nebius.com]

[36] EuroHPC JU (2024). ***The EuroHPC Joint Undertaking Launches AI Factories Calls to Boost European Leadership in Trustworthy AI*** [eurohpc-ju.europa.eu]

[37] TNO (2024). ***Quickscan AI in de Publieke Dienstverlening III*** [publications.tno.nl]

[38] Europese Commissie: DG Research & Innovation (2024). ***Align, act, accelerate – Research, technology and innovation to boost European competitiveness*** [data.europa.eu] Publications Office of the European Union.

dialogic

Onderzoek voor *onderbouwd* beleid.

Dialogic innovatie & interactie
Hooghiemstraplein 33
3514 AX Utrecht
030-2150580
www.dialogic.nl