

APPLICATIONS UNLEASHED 2024

The era of
prompt innovation



TABLE OF CONTENTS

		01	02
	VOORWOORD	DE INVLOED VAN AI OP ORGANISATIES	CHAT IS DE NIEUWE SUPER APP
	02	03	06
03	04	05	06
KICKSTART AI MET MANAGED AI IN DE CLOUD	DE IMPACT VAN HYPERSCALERS OP GEN AI	AI ALS POWER BOOSTER VOOR ERP	DE AI-EVOLUTIE IN SOFTWARE TESTING
10	15	18	22
07	08	09	10
WORDT EEN LARGE LANGUAGE MODEL STRAKS JOUW NIEUWE COLLEGA?	VAN IDEE TOT SOFTWARE MET GENERATIEVE AI ALS GAMECHANGER	SUSTAINABILITY BY DESIGN: EEN CRUCIALE NIEUWE NORM IN CLOUD-ARCHITECTUUR	MISSION IMPOSSIBLE: AUTOMATIC USABILITY TESTING
26	31	36	41
11	12	13	
PRODUCT DISCOVERY: ECHTE WAARDE CREËREN IN PLAATS VAN ONGEBRUIKTE FEATURES BOUWEN	PLANT SIMULATION UNLEASHED: DE SLEUTEL TOT MANUFACTURING EXCELLENCE EN IT-SUPPORT	ROBUUSTHEID VAN SOFTWARE ENGINEERING DOOR MIDDEL VAN HYPER-AUTOMATISERING	COLOFON
45	50	55	61

VOORWOORD

Beste lezer,

Dankzij technologie is de mens tot meer in staat dan we ooit hadden kunnen dromen. Technologie maakt ons creatiever, effectiever, communicatiever, innovatiever. Maar er is ook een keerzijde; technologie zet ons niet alleen in onze kracht, ze beïnvloedt steeds meer ons denken en handelen, is steeds bepalender voor de manier waarop we ons organiseren, en is doorgedrongen tot in de haarvaten van onze maatschappij. En: als onze opvattingen en ideeën over de rollen, taken en ontwikkeling van technologie tekortschieten, zijn de technologische uitkomsten onzeker – en daarmee de uitkomsten voor de mens. Zo ontstaat een steeds grotere wederzijdse afhankelijkheid, tussen mens en technologie. De kunst is dat alles in goede banen te leiden. Een toekomst vorm te geven waarin wij als mens aan het roer blijven, terwijl technologie ons helpt de juiste kant op te navigeren. Het is aan ons de juiste, menselijke maat te bewaken in input en output, in de wetenschap dat technologisch gezien de mogelijkheden inmiddels echt eindeloos zijn. Dit, in het kort, is de context waarbinnen deze nieuwe editie van Application Unleashed zich beweegt. De titel: The Era of Prompt Innovation.

Application Unleashed kunt u beschouwen als een 'bijsluiter' van Capgemini's overkoepelende TechnoVision: augment ME! Onze visie reikt ver voorbij traditionele IT-trends; wij definiëren een breed spectrum van technologieën die samen aan de bron staan van de huidige, wereldwijde, allesomvattende metamorfose. Vanuit TechnoVision willen we de strategische dialoog aanjagen tussen technologen en business leaders, om zo uit te komen op de juiste prioriteiten voor een duurzame digitale toekomst.

De voornaamste trigger voor dit alles is natuurlijk Generatieve AI. Sinds deze technologie in 2023 grootschalig haar intrede deed, zijn we continu bezig ons denken bij te stellen over wat technologisch mogelijk en wenselijk is. Het is dus geen wonder dat TechnoVision GenAI ziet als een van de 5 'megatrends' van dit moment. En 2024 wordt een cruciaal jaar; zijn we in staat de verdere ontwikkeling van AI – en de vormgeving van onze maatschappij rond dit thema – de juiste kant op te sturen?

Maar er is veel meer gaande. Denk aan Quantumtechnologie, halfgeleiders, accu-technologie, SpaceTech... Niet eerder gebeurde er zoveel op technologisch gebied, en was de potentiële impact zo groot. Application Unleashed fungeert als een leidraad bij al deze ontwikkelingen.

Daarbij kiezen we een pragmatische insteek. Want of we het nu hebben over sustainability by design, de impact van hyperscalers op het AI-landschap, of de razendsnelle

Application Unleashed kunt u beschouwen als een 'bijsluiter' van Capgemini's overkoepelende TechnoVision: augment ME!

ontwikkeling van chat-platforms: het zijn concrete ontwikkelingen met concrete gevolgen. Niet alleen puur technologisch, maar vooral ook voor de manier waarop wij als mens ons tot technologie verhouden. Iedereen die erbij betrokken is moet zich daarvan bewust zijn, er bovenop zitten, en een strategische opvatting hebben over de volgende stappen.

Welke onderwerpen kunt u dan verwachten? We bespreken de broodnodige integratie van duurzaamheid in cloud-architectuur; er liggen hier grote kansen, op weg naar een groenere toekomst. We onderzoeken het kritieke belang van geautomatiseerde usability testing; niet alleen vanuit technologisch perspectief, maar ook vanuit het perspectief van de menselijke ervaring. We gaan dieper in op de status van chat als de nieuwe 'super app'; dat is niet alleen een technologische aardverschuiving, maar heeft ook grote gevolgen voor hoe wij als mens communiceren met technologie en met elkaar. En er is meer, veel meer.

In deze tijd waarin prompting steeds meer de basis vormt voor onze interactie met technologie, moeten we er samen voor zorgen dat onze 'prompts' de juiste zijn. Zodat we samen vormgeven aan een duurzame, technologie-gedreven toekomst – met behoud van de menselijke maat, en onze menselijke waarden. De mogelijkheden zijn eindeloos; aan ons de taak het juiste mogelijk te maken. Application Unleashed – The Era of Prompt Innovation biedt u hierbij een concrete roadmap.

Veel leesplezier!



DR. CARA ANTOINE

Executive Vice President,
Chief Technology
& Innovation Officer,
Capgemini Europe

01

DE INVLOED VAN AI OP ORGANISATIES

EEN INTRODUCTIE IN KUNSTMATIGE INTELLIGENTIE

Werken in organisaties is constant aan verandering onderhevig. Aangedreven door innovatie, veranderende klantbehoeften en het streven naar efficiëntie, lijkt het alsof dat de enige constante waarde is in de organisatie van nu. De veranderingen kunnen gepaard gaan met angst, zoals de verwachting dat de introductie van de computer banen zou kosten. Ondanks dat deze verwachting niet juist was¹ veranderde de komst van computers het werk.

Nu gaat door de introductie van Artificiële Intelligentie (AI) werk opnieuw veranderen. Bij het brede publiek wordt bij AI direct gedacht aan ChatGPT. Toch waren daarvoor ook al toepassingen voorhanden die gebruik maken van AI: zo werd schaakgrootmeester Kasparov in 1997 verslagen door de IBM-computer Deep Blue² en kan Google Translate³ sinds 2006 worden gebruikt om teksten te vertalen.

DEFINITIE AI

Maar wat is AI nou eigenlijk?

John McCarthy van Stanford University gaf in 2007 de volgende definitie aan AI: het is de wetenschap voor het maken van intelligente machines en software⁴. De opkomst van AI is versneld door de beschikbaarheid van zeer grote hoeveelheden data en de mogelijkheid deze data te analyseren en modeleren in serverparken. De essentie van alle AI-modellen is zo accuraat mogelijk de uitkomst te voorspelen op basis van de input⁵.



1. Boom Management, 2023
2. MIT Technology Review, 2022
3. Sommerlad, 2021
4. McCarthy, 2007
5. Open Universiteit, 2020

WAT KAN AI NU AL VERANDEREN?

AI wordt gebruikt als een containerbegrip en is op te delen in drie subgroepen:

- **Artificial Narrow Intelligence (ANI):** AI dat specifiek getraind is voor een applicatie en met een bepaald doel voor ogen
- **Artificial General Intelligence (AGI):** AI dat gemaakt is om meerdere algemene taken te kunnen verrichten zonder hier specifiek op getraind te zijn
- **Artificial Super Intelligence (ASI):** AI dat de meest generieke taken beter kan uitvoeren dan mensen

Voor de twee laatstgenoemden bestaan er geen praktische voorbeelden en is dit puur theoretisch van aard (IBM, sd). Op dit moment is ANI nog geen volwaardige tegenhanger van menselijke intelligentie. Een oorzaak hiervan is dat de getrainde modellen het vermogen missen om onvoorziene of nieuwe scenario's op te pakken iets wat het menselijk brein zo goed kan. Bijvoorbeeld, een mens met een rijbewijs kan in allerlei merken auto's rijden, terwijl een AI-toepassing zich in één merk heeft getraind.

Op dit moment heeft ANI dus de meeste impact op de manier hoe we werken. Hieronder vallen verschillende technieken en mogelijke toepassingen. De technieken die binnen softwareontwikkeling vaak gebruikt worden zijn Machine Learning (ML), Neural Network (NN) en Deep Learning (DL). Deze technieken worden vaak door elkaar gebruikt, maar het is beter om deze te beschouwen als een Matroesjka pop. Dit zijn in elkaar passende houten poppetjes.

De grootste Matroesjka pop staat voor Machine Learning. Twee voorbeelden hiervan zijn:

- In de financiële dienstverlening kunnen fraudepatronen herkend worden door ML. De software krijgt bij de ontwikkeling gelabelde data van correcte en frauduleuze transacties om van te leren. Op basis hiervan kan het systeem voorspellen of een transactie frauduleus is of niet. Dit is een voorbeeld van supervised Machine Learning (Capgemini, 2023).
- Een tweede voorbeeld van supervised Machine Learning is beeldherkenning: denk hierbij aan een producent van koffiemokken, waarbij de machine een foto van het eindproduct (de koffiemok) maakt en vergelijkt met het beeld. Dit kan het systeem doen doordat het getraind is op basis van gelabelde data. Op voorhand is dus aan het systeem getoond hoe een goede koffiemok eruit moet zien.

Het middelste Matroesjka poppetje is een Neural Network (NN). De opzet kan vergeleken worden met de neuronen in het menselijke brein. Deze techniek biedt de mogelijkheid om ongelabelde data te analyseren. Google heeft in een experiment een NN een week lang video's van YouTube laten analyseren. Uiteindelijk herkende het NN op de beelden of er een kat op staat of niet zonder dat het systeem uitleg heeft gekregen over wat een kat is⁶.

De input voor de NN in dit voorbeeld zijn YouTube video's. Vervolgens wordt de analyse gedaan door de verborgen lagen die met elkaar zijn verbonden zijn en uiteindelijk geeft de outputlaag de waarschijnlijkheid aan of er op het beeld een kat staat of niet.

Het kleinste Matroesjka poppetje is Deep Learning. Dit is vergelijkbaar met een Neural Network, maar het verschil is dat dit netwerk meer dan drie tussenlagen heeft (IBM, sd). DL kan gebruikt worden om teksten, beelden of geluid te genereren, zoals de fotos van de Matroesjka pop. Een bekender voorbeeld is een video op YouTube van Queen Elizabeth II die het in haar kersttoespraak heeft over TikTok en een dansje oefent. Channel 4 in de UK heeft deze video gemaakt om de mogelijkheden en risico's van deze techniek te tonen⁷.

WAT IS DE IMPACT VAN DEZE VERANDERING?

De beschreven AI-technieken zullen onze manier van werken gaan veranderen. Het biedt de mogelijkheid om verschillende repetitieve taken en analyses uit te laten voeren. Deze taken kunnen Artificial Narrow Intelligence toepassingen beter uitvoeren dan de mens, omdat ze getraind zijn op extreem grote datasets⁸.

Het bekendste voorbeeld is waarschijnlijk ChatGPT. Dit is een chatbot waarbij GPT staat voor Generative Pre-trained Transformer. Andere termen die gehanteerd worden zijn Generatieve-AI of Large Languages Model. Ondanks dat deze systemen slim lijken komt het niet in de buurt van de eerder vermelde AGI. Op hoofdlijnen werken de toepassingen als volgt. Het systeem heeft een zeer grote dataset, in het geval van bijvoorbeeld Llama (chatbot van Facebook), 2000 miljard tokens (Meta Platforms, 2023). Hierbij is een token een stukje van een woord. Als de gebruiker een vraag stelt probeert het systeem op basis van alle beschikbare data te voorspellen welk antwoord zo dicht mogelijk bij de vraag komt⁹. Toch brengt generatieve AI ook risico's en heeft het tekortkomingen. Denk hierbij aan verschillen tussen versies van het model, ChatGPT 3 of 4. Ook kan het voorkomen dat het systeem gaat hallucineren. Dit betekent dat het model een output geeft die niet overeenkomt met de trainingsdata.

6. Google, 2012

7. Channel 4 Entertainment, 2020

8. Banafa, 2023

9. MIT, 2023

Het bekendste voorbeeld is waarschijnlijk ChatGPT. Dit is een chatbot waarbij GPT staat voor Generative Pre-trained Transformer. Andere termen die gehanteerd worden zijn Generatieve-AI of Large Languages Model.

Fotos: Matroeskja poppen gegenereerd door Adobe Firefly, een generatieve AI oplossing.



VERSCHILLENDE KANTEN VAN AI

In deze editie van Applications Unleashed belichten we verschillende kanten van AI.

- **Ron Tolido** schrijft over natuurlijke taal als uniforme toegang tot applicaties.
- **Pascal Regeer** gaat uitgebreid in op de mogelijkheden en tekortkomingen van generatieve AI.
- Het analyseren van extreem grote datasets zoals die gebruikt worden binnen Llama en ChatGPT kan alleen efficiënt worden gedaan in serverparken. Deze partijen worden hyperscalers genoemd. Er zijn verschillende commerciële aanbieders van zulke oplossingen zoals bijvoorbeeld AWS, Google en Microsoft. Lees het artikel van **Sjoukje Zaal** om meer te weten over dit onderwerp.
- **Malin de Silva** beschrijft verschillende toepassingen van Managed AI in de cloud.
- Het ontwikkelen van software wordt een creatiever beroep door de inzet van een digitale assistent bij het programmeren. Lees hiervoor het artikel van **Donald Hessing**.
- Moet software nog wel door mensen getest worden? **Sanne Visser** en **Menno Bruijn** leggen uit waarom de tester nodig blijft en misschien nog wel belangrijker wordt.
- Het ERP systeem is het digitale hart van alle grote organisaties, waarin alle bedrijfsdata zijn opgeslagen. Deze beschikbaarheid van informatie en rekencapaciteit op de server leent zich bij uitstek voor verschillende analyses. Lees de toekomstvisie op ERP en AI in het stuk van **Zahir Wanga** en **Marc Steenberg**.

OVER DE AUTEUR



RIKKERT WIJRDEMAN

Test Automation Engineer

Rikkert Wijrdeman is een IT-consultant met meer dan 5 jaar werkervaring in testautomatisering. Hij heeft verschillende rollen vervuld, waaronder Auditor, Testcoördinator en Testautomatiseerder. Zijn nieuwsgierigheid naar waarom processen lopen zoals ze lopen, en hoe ze verbeterd kunnen worden, is iets wat Rikkert typeert. Rikkert is daarnaast expert op het gebied van kunstmatige intelligentie.

[in https://www.linkedin.com/in/rikkert-w-86b99925/](https://www.linkedin.com/in/rikkert-w-86b99925/)

✉ rikkert.wijrdeman@capgemini.com



02

CHAT IS DE NIEUWE SUPER APP

**NATUURLIJKE TAAL ALS UNIFORME
TOEGANG TOT APPLICATIONS**

HIGHLIGHTS

- De Super App is het Zwitserse zakmes van mobiele applicaties: één applicatie die al je taken uitvoert, zonder dat je hoeft te switchen van applicatie.
- Wie een leidende Super App weet te maken, krijgt alle touwtjes in handen rond gebruikersinformatie en applicatiediensten.
- Met de grootste voorbeelden allang functioneel in Azië, richt ook de rest van de wereld zich op het maken van een leidende Super App.
- Met de opkomst van ChatGPT en open plug-ins ontwikkelt de dialoog in natuurlijke taal zich onverwacht tot zo'n Super App.
- Ontwikkelaars zouden hierop moeten anticiperen, door hun applicatiediensten beschikbaar te stellen als plug-ins.

Elon Musk zou enkel en alleen hierom Twitter hebben gekocht. Het is de droombaby van Gartner. In China lusten ze er wel pap van. De Super App is de 'one ring to rule them all' van de applicatiewereld. Er is veel concurrentie, maar op de achtergrond wacht een onverwachte speler rustig af: natuurlijke taal.

Welkom bij de Super App, het Zwitserse zakmes van mobiele applicaties: één applicatie die al je taken uitvoert, zonder dat je hoeft te switchen van applicatie, zonder dat je steeds weer met een andere gebruikersinterface krijgt te maken. Wie een leidende Super App weet te maken, krijgt alle touwtjes in handen rond gebruikersinformatie en applicatiediensten. Er is veel discussie over welk mobiel platform tot zo'n Alpha App zou kunnen uitgroeien en leveranciers van applicaties volgen de ontwikkelingen met argusogen. Maar in de coulissen staat een hoogst onverwachte speler koeltjes af te wachten om de hoofdrol over te nemen. Wat is toegankelijker en uniformer dan natuurlijke taal; het woord zegt het immers al? Met de doorbraak van ChatGPT, gecombineerd met de opkomst van op open standaards gebaseerde plug-ins, zou een dialoog in natuurlijke taal wel eens de definitieve Super App kunnen blijken te zijn.

HET ONTSTAAN VAN SUPER APPS

Het begrip 'Super App' staat op het conto van Mike Lazaridis, de oprichter van het notoire bedrijf Research in Motion (RIM), de ontwikkelaar en producent van de toentertijd fameuze BlackBerry-apparaten. Helemaal terug in 2010 definieerde hij ze al als een gesloten ecosysteem van talloze applicatiediensten die mensen elke dag zouden gebruiken, waarbij het enthousiasme zou voortkomen uit een naadloze, totaal geïntegreerde, en contextgevoelige gebruikerservaring. Geen gedoe met het vinden van de juiste applicatie of het moeten leren van een nieuwe gebruikersinterface: de Super App biedt toegang tot alles, biedt een uniforme manier van gebruik en – heel belangrijk – onthoudt de context van de gebruiker van de Super App. Op die manier hoeft de gebruiker niet steeds weer alle benodigde informatie in te voeren.

Toegegeven, de visie van Lazaridis heeft het uiteindelijk niet helemaal gemaakt via de BlackBerry. Waar ze er wel mee aan de haal zijn gegaan is in Azië, waar platforms zoals WeChat (in China) en Grab (in Zuidoost-Azië) compleet ingeburgerd zijn geraakt. Ze worden beschouwd als de vroege voorbeelden van echte Super Apps. Deze applicaties bieden een breed scala aan diensten, zoals berichten, e-commerce, transport en mobiele betalingen, allemaal onder één dak, waardoor gebruikers niet de aandrang krijgen om hun heil bij andere applicaties te zoeken.

Super Apps beginnen vaak met één eenvoudige basisfunctie en stapelen daar geleidelijk aanvullende oplossingen en diensten op – uiteraard gedreven door synergie en gemeenschappelijkheid. Het enorme Chinese technologieconglomeraat Tencent bijvoorbeeld, transformeerde zijn 'WeChat' app geleidelijk naar een echte Super App. Wat begon als een berichtenapplicatie veranderde in een ecosysteem van diensten, waaronder het boeken van taxiritten, virtuele portefeuilles, het maken van hotelreserveringen, het aanvragen van medische consultaties en meer. Grab begon in Maleisië met het aanbieden van vervoer, maar evolueerde naar een Super App waarmee elke vorm van transport kan worden geregeld, aangevuld met allerlei financiële en zakelijke diensten – en zelfs videostreaming.

DE SLAG OM DE SUPER APP

Een Super App werkt als een platform: naarmate zowel vraag als aanbod blijven toenemen wordt de Super App aantrekkelijker en bekender – met als resultaat nog meer vraag een aanbod. Het is voor de leverancier van het platform zaak de exclusieve aandacht te trekken van de ontwikkelaars van applicatiediensten. Dat kan gebeuren door het aanbieden van zoveel mogelijk middelen om een applicatie te ontwikkelen, in de vorm van API's, applicatieraamwerken, goed gedefinieerde standaards en zelfs ondersteuning voor het ontwikkelen van code. Combineer het met hebben van de snelst groeiende gebruikerspopulatie en het al aanbieden van andere diensten die synergetisch potentieel hebben, en het Super App platform werkt als een pak suiker op een mierenkolonie.

Waar de Super App tot nu toe vooral in Azië is doorgebroken (let bijvoorbeeld ook op de volgende stappen van TikTok), neemt dat niet weg dat er elders spelers hun pijlen hebben gericht op het maken en leveren van een Super app. Amazon bijvoorbeeld – oorspronkelijk bekend als de leverancier van een website voor boeken – heeft zich niet alleen ontwikkeld tot een leverancier van een brede waaier aan e-commerce-diensten, maar biedt nu ook bijvoorbeeld financiële diensten (Amazon Pay), en entertainment (Amazon Prime). Die gaan vroeg of laat samenkomen in iets wat verdacht veel op een Super App zal lijken. Het populaire Snapchat heeft een stroom van additionele diensten gelanceerd, uitwaaiend naar nieuwe markten. Meta is WhatsApp aan het uitbreiden met financiële functionaliteiten, en wie weet wat straks nog meer. En dan is daar Elon Musk die – ongetwijfeld met de successen van zijn vroegere geesteskind PayPal in het achterhoofd – grote plannen heeft om zijn 'X' (TAFKAT, The App Formerly Known As Twitter) tot een allesomvattende Super App uit te laten groeien.

Kortom, machtige spelers draaien rondjes om elkaar op zoek naar de verleiding van zowel de klanten als de ontwikkelaars en leveranciers van applicatiediensten, op zoek naar een positie als de leidende producent en eigenaar van een Super App. Mobiele toestellen en bijbehorende gebruikersinterfaces lijken daarbij een gegeven te zijn, als het meest voor de hand liggende 'Super Channel' (VR en AR laten we voorlopig nog maar even buiten beschouwing).

DE SUPERKRACHT VAN TAAL EN CONVERSATIE

Hoog tijd om een blik werpen op een onverwachte nieuwkomer in dit onvoorspelbare schouwspel: ChatGPT, het vernuftige natuurlijke taal en conversatie AI-systeem van OpenAI. In toenemende mate nu gecombineerd met spraaktechnologie heeft deze applicatie (voor zover dat nog een accurate term is) het potentieel om uit te groeien tot een dominante Super App – of ‘Super App 2.0’ voor degenen die nog steeds gek zijn op versienummers).

Om te begrijpen waarom AI chat (disclaimer: dit betoog heeft evenzeer betrekking op concurrenten van ChatGPT, zoals Bard van Google) het potentieel heeft om een Super App te worden, moeten we eerst de kracht van AI voor conversatie en natuurlijke taal onder de loep nemen. ChatGPT en vergelijkbare systemen zijn gebouwd op geavanceerde machine learning-technologie en is getraind op enorme hoeveelheden tekst uit het internet. Dit heeft geresulteerd in een AI-model dat in staat is om natuurlijke menselijke taal te begrijpen en te genereren. Gecombineerd met Reinforcement Learning is daar ook een spectaculaire functionaliteit aan toegevoegd in termen van het stapsgewijs kunnen voeren van een dialoog, met behoud van context, om de uiteindelijke intentie helder te krijgen.

Dit vermogen om te communiceren in natuurlijke taal en een dialoog aan te gaan, is baanbrekend. Het betekent dat AI chat niet alleen informatie kan verstrekken, maar ook kan begrijpen wat de gebruiker wil en kan reageren met menselijke context. Dit maakt de interactie veel intuïtiever en flexibeler dan traditionele mobiele interfaces. Er is immers een goede reden waarom we ‘natuurlijke taal’ zo noemen: dichterbij dan taal komen we niet als we effectief en doelgericht willen communiceren. Toepassingen zoals ChatGPT zijn daarom ondertussen in een opmerkelijk korte periode de favoriete plek geworden van velen, niet alleen om taal en plaatjes te generen, maar ook antwoorden te vinden en – waarom niet – om zaken gedaan te krijgen.

Amazon bijvoorbeeld – oorspronkelijk bekend als de leverancier van een website voor boeken – heeft zich niet alleen ontwikkeld tot een leverancier van een brede waaier aan e-commerce-diensten, maar biedt nu ook bijvoorbeeld financiële diensten (Amazon Pay), en entertainment (Amazon Prime).

PLUG-INS: DE BOUWSTENEN VAN DE AI CHAT SUPER APP

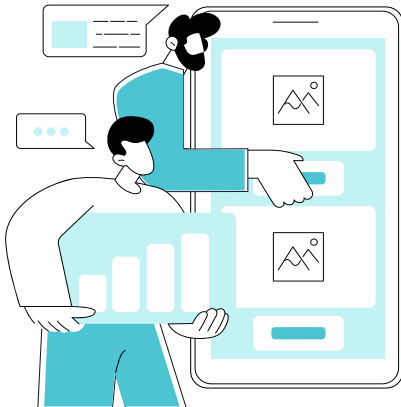
Voor dat laatste hebben we een speciale, additionele superkracht nodig: de plug-ins. Deze kleine toegevoegde componenten, bijvoorbeeld geschikt voor ChatGPT, verbinden AI chat met applicatiediensten van derden. Ze stellen het AI-systeem in staat om te communiceren met API's die door externe ontwikkelaars zijn ingevuld en geleverd, waardoor de mogelijkheden van AI chat aanzienlijk worden uitgebreid. Tegelijkertijd worden met API's vangrails aangeboden om de consistentie, veiligheid en juistheid van zowel informatie als acties te kunnen genereren.

Plug-in-ontwikkelaars voor ChatGPT werken op basis van OpenAPI-specificatie. Deze is volledig op meta-definities gebaseerd, waardoor er geen beperking is aan het soort API wat uiteindelijk aan ChatGPT wordt aangeleverd. Het ontwikkelen komt neer op het bouwen van een API (en de applicatiediensten erachter natuurlijk), het vastleggen van de API in het OpenAPI formaat en het definiëren van de metadata van de API in een ‘manifest’ bestand. Dit alles wordt vervolgens geregistreerd en vastgelegd voor gebruik vanuit een ChatGPT ‘plug-in store’.

Hoewel voor ChatGPT de plug-ins alleen nog in bèta-versie beschikbaar zijn, is er al een hele lijst van al dan

niet veelbelovende toevoegingen – velen met de allure van een functionaliteit die je in een Super App zou willen zien, bijvoorbeeld:

- **OpenTable**, voor het reserveren van een restaurant
- **Expedia** en Kayak, voor het regelen van reizen en verblijven
- **Instacart**, voor recepten en ingrediënten
- **Chess**, voor het spelen van schaak
- **World Bank Data**, voor toegang tot de brede World Bank datasets
- **Ask Your PDF**, voor het hebben van een 'conversatie' met een PDF-document
- **What To Watch**, voor het vinden van streaming programma's en series



De primaire toegang tot een klant zou straks weleens niet meer een - al dan niet mobiele - applicatie, kunnen zijn, maar een kleine plug-in voor een AI Chat platform.

Veel meer gaat komen, zeker als de plug-in functionaliteit breder beschikbaar wordt en meer ervaring is opgedaan met de betrouwbaarheid en accuratesse van plug-ins.

OP MACH-SPEED VERDER

Bedrijven – en hun applicatieontwikkelaars – doen er goed aan om de verdere evolutie van AI Chat en plug-ins op de voet te volgen. Het denken vanuit de moderne principes van 'MACH' (Microservices, API-first, Cloud-native, Headless) wordt er in ieder geval alleen maar door gescherpt, want dat is de moderne architectuur om applicatiediensten beschikbaar te maken voor AI Chat. De primaire toegang tot een klant zou straks weleens niet meer een – al dan niet mobiele – applicatie, kunnen zijn, maar een kleine plug-in voor een AI Chat platform. De beste applicatie is misschien wel geen applicatie. En de beste website zou er straks weleens een stuk leger kunnen uitzien.

Koppel al dit fraais met systemen voor spraakherkenning en -generatie (o hallo Alexa, Siri, weer terug?), en het wordt duidelijk dat Elon Musk misschien al die tijd al het materiaal van een Super App in handen had. Het was alleen niet X. Het is de spraakherkenning in zijn Tesla's.

OVER DE AUTEUR



RON TOLIDO

CTO, EVP, Master Architect

Ron is de Chief Technology Officer van Capgemini's Insights & Data practice. Hij richt zich op de impact van baanbrekende innovaties op de bedrijfsvoering. Ron is de hoofdauteur van Capgemini's jaarlijkse 'TechnoVision' trendrapport, de bedenker van de term 'Applications Unleashed' en een frequente gastdocent op Universiteiten en Hogescholen.

<https://www.linkedin.com/in/rtolido/>

ron.tolido@capgemini.com

03

KICKSTART AI MET MANAGED AI IN DE CLOUD

EEN ANALYSE VAN DE OPTIES, DE VOORDELEN EN DE UITDAGINGEN VAN MANAGED AI SERVICES IN DE CLOUD

HIGHLIGHTS

- Het faciliteren van Managed AI-diensten is een belangrijke prioriteit voor cloud-aanbieders.
- Alle vooraanstaande cloud-aanbieders bieden vergelijkbare diensten, die ook nog eens eenvoudig zijn toe te passen.
- Het gemak van onboarding in de cloud, beprijzing op basis van gebruik, lage onderhoudslasten en betere SLA's zijn goede redenen om te kiezen voor Managed AI in de cloud.
- Beginnen met Managed AI in de cloud is eenvoudig, maar om de bruikbaarheid en effectiviteit voor de lange termijn te garanderen is een goed gedefinieerd target operating model onontbeerlijk.
- Je moet een strategische visie ontwikkelen op de scope en de roadmap; je kunt immers voor uitdagingen worden gesteld die de capaciteiten van Managed AI te boven gaan.

Managed Artificial Intelligence services in de cloud kunnen de ontwikkeling van AI-applicaties flink versnellen. Er zijn verschillende opties beschikbaar, afkomstig van allerlei cloud-dienstverleners en ze zijn eenvoudig in het gebruik. In dit artikel gaan we dieper in op de diensten, de mogelijkheden en de overwegingen.

Heb je een klein of middelgroot bedrijf? Zou je een gepersonaliseerde chatbot willen inrichten, om je klanten meer te betrekken bij je merk, maar heb je geen idee hoe je dat aanpakt? Of heb je een groot bedrijf dat al een enorme hoeveelheid klantdata en feedback heeft vergaard, maar heb je geen idee hoe je daarin prioriteit aanbrengt? Klantenservice is een belangrijke use case voor data-gedreven besluitvorming; de data-gedreven praktijk blijkt echter vaak weerbarstig. Effectief gebruik van data voor klantdoeleinden vereist flinke investeringen en veel kennis van zaken.

Tenminste, zo was het altijd, voordat Managed AI-dienstverleners hun intrede deden. De toepassing van AI is een complex proces dat bestaat uit verschillende fasen, uiteenlopend van probleemdefiniëring tot dataset-preparatie en nog veel meer. En elke fase vereist specifieke skills en investeringen. Managed AI-diensten in de cloud werken hier drempelverlagend en bieden van tevoren getrainde datamodellen die je met een relatief bescheiden inspanning kunt finetunen voor toepassing in de klantbenadering van jouw bedrijf.

De cloud-based oplossingen van Managed AI-dienstverleners worden steeds populairder; in de vorm van API's kunnen ze eenvoudig geïntegreerd worden met moderne applicaties.

WAT IS MANAGED AI IN DE CLOUD?

Managed Artificial Intelligence-diensten vormen een integraal onderdeel van de diensten die worden aangeboden door alle vooraanstaande cloud-platformen. De meeste diensten worden aangeboden als PaaS-oplossingen; de aanbieder beheert in zo'n structuur alles, van infrastructuur tot en met de pre-trained AI-modellen. De klant kan het model aanpassen, valideren en optimaliseren voor toepassing binnen de eigen omgeving. De pre-trained modellen worden vaak door de cloud-dienstverlener onderhouden en geüpdate; de klant ondervindt daar nauwelijks hinder van. Afgezien van de eigen diensten bieden cloud-platforms ook marktplaatsen waar andere aanbieders hun eigen Managed AI-producten kunnen aanbieden.

Het is eenvoudig om met Managed AI aan de slag te gaan en verschillende diensten aan te sluiten. Tekst destilleren uit een document kan bijvoorbeeld door het document via een REST API up te loaden in de managed service: de managed service haalt vervolgens de tekst uit het document en stuurt die naar je terug. Op dat moment kun je bijvoorbeeld een vertaaldienst gebruiken om de content te vertalen in de gewenste taal. En zo zijn er nog veel meer diensten beschikbaar, die allemaal een ander soort output kunnen genereren.

De managed AI-diensten in de cloud kunnen op dezelfde locatie gehost worden als de applicaties waar de workloads uitgevoerd worden. Dat zorgt voor lage latency en hoge efficiency.

DE VERSCHILLENDE TYPEN BESCHIKBARE DIENSTEN

Er worden verschillende soorten Managed AI-diensten aangeboden, in de vorm van customizable PaaS-oplossingen

Zoekmachines

Moderne zoekmachines zijn gepersonaliseerd. Veel van hun eigenschappen zijn inmiddels beschikbaar in de vorm van Managed AI-diensten. Ze kunnen worden gebruikt om informatie in te winnen binnen een bepaalde context, zoals een organisatie. Je kunt je eigen zoekindexen creëren op basis van data die beschikbaar zijn binnen de organisatie, zodat gebruikers snel de informatie krijgen waar ze om vragen in de zoekmachine.

De vooraanstaande cloud-dienstverleners bieden de volgende zoekmachines aan:

Microsoft Azure	Azure Cognitive Search
Amazon Web Services	Amazon Kendra
Google Cloud	Vertex AI Search



Taaldiensten

Dankzij AI zijn tegenwoordig geavanceerde taaldiensten beschikbaar in de cloud. Moderne chatbots begrijpen wat je van ze vraagt, en kunnen betekenisvolle antwoorden bieden, op basis van een steeds groter volume aan data en informatie. Er bestaan ook spraakdiensten die gesproken woorden begrijpen. Spraakdiensten zijn van grote waarde bij het bieden van een pakkende klantervaring.

Er zijn ook verschillende, tekstgebaseerde analysefuncties beschikbaar. Een voorbeeld is entiteit-herkenning. Die functie is in staat om terugkerende elementen in teksten te identificeren – namen van mensen bijvoorbeeld. Een ander voorbeeld is het samenvatten van lange teksten. Die functie kan van grote waarde zijn voor specifieke use cases. We onderscheiden hier twee typen:

- Extractive samenvattingen halen de voornaamste conclusies uit een document
- Abstractive samenvattingen brengen de volledige tekst terug tot bescheidener proporties, met behoud van de betekenis

De voornaamste cloud-aanbieders bieden de volgende taaldiensten aan:

	Chatbots & Virtual Agents	Text to Speech	Speech to Text
	Azure AI Bot Service	Azure AI Speech	Azure AI Speech
	Amazon Lex	Amazon Polly	Amazon Transcribe
	Dialogflow	Amazon Transcribe	Speech-to-Text

Computer Vision

Beeld, en video, analyse zijn belangrijke diensten in vele toepassingen. Dergelijke diensten bieden ondersteuning bij de identificatie van objecten en mensen en kunnen zelfs kenmerken classificeren zoals emoties, gelaatsuitdrukkingen en hulpmiddelen die afgebeelde mensen dragen, zoals brillen. Dergelijke analyses kunnen ook worden gebruikt om inspecties van assemblagelijnen te faciliteren. De image processing die hiervoor nodig is kan veel bandbreedte vergen; dat levert soms praktische uitdagingen rond connectiviteit op. De effectiviteit van video-analyse kan immers worden beïnvloed door de kwaliteit van de video en het aantal frames.

De voornaamste cloud-aanbieders bieden de volgende computer vision-diensten aan:

	Text Extraction from Images (OCR)	Face Recognition	Object Detection and Analysis
	Optical Character Recognition in Azure Vision Studio	Face in Azure Vision Studio	Spatial Analytics in Azure Vision Studio
	Amazon Textract	Amazon Rekognition	Amazon Lookout for Vision
	OCR in Cloud Vision API	Face detection (geen gezichtsherkenning)	Cloud Vision API



Machine Learning

Machine Learning gebruikt data en algoritmen om een oplossing te bieden voor menselijke beperkingen in besluitvorming, waarbij ML het menselijke leerproces imiteert. Machine Learning in de cloud is een complexe aangelegenheid waarvoor data engineers en wetenschappers nodig zijn. Alle vooraanstaande cloud-aanbieders ondersteunen een efficiënter machine learning-proces, met drag and drop-tooling en geautomatiseerde leer-hulpmiddelen.

Generative AI use cases gooien momenteel hoge ogen. Ook deze worden vaak gebouwd en geoptimaliseerd aan de hand van ML-modellen die onder menselijke supervisie staan en die zijn toegesneden op de use case.

Naast deze kern-capaciteiten en diensten bestaan er nog veel meer Managed AI-diensten. De ontwikkelingen gaan bovendien razendsnel.

FROM ZERO TO HERO

Strategie

Het onboarding-proces van Artificial Intelligence binnen organisaties kan enorm ingewikkeld zijn. Het is daarom aan te raden om van tevoren het juiste Target Operating Model vast te stellen. De meeste data-gedreven strategieën bestaan uit een samenspel tussen People, Process en Technology, maar er kunnen – afhankelijk van de toepassing en de sector waarin de organisatie opereert – nog meer kritische aspecten bestaan.

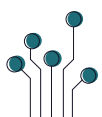
Binnen de aspecten van People, Process en Technology zijn de volgende strategische overwegingen van belang:



- **People:** Rollen die zich bezighouden met AI moeten kennis en ervaring hebben met data-engineering en architectuur. Daarnaast moeten ze vendor-specifieke cloudvaardigheden hebben. Bovendien zullen bepaalde governance-rollen nodig zijn.



- **Processes:** De inbedding van AI in de organisatie omvat verschillende processen, zoals data capturing, filtering, validation, finetuning en go-live. Deze moeten van tevoren worden vastgesteld en de juiste mensen moeten worden toegewezen aan het beheer en de uitrol ervan.



- **Technology:** Als je al een bepaalde cloud-aanbieder hebt gekozen, is het logisch om deze ook te selecteren als aanbieder van Managed AI-diensten. Sommige organisaties hebben verschillende cloud-aanbieders aan boord, en daarmee verschillende alternatieven voor dezelfde capabilities. Als je nog geen aanbieder hebt uitgekozen, is het raadzaam om eerst de markt te verkennen."

IMPLEMENTATIE

Zolang de backlog van de organisatie het toelaat, kost het maar een paar minuten om een Managed AI-dienst binnen de cloudomgeving van de klant te implementeren. Het is daarom heel goed te doen om beperkte use cases te testen, nog voordat je je proof of concept helder voor ogen hebt.

Je eerste PoC kan iets eenvoudigs zijn als een chatbot die vragen van klanten over je producten beantwoordt. Mocht je Managed AI willen gebruiken in meer business-kritische toepassingen, dan is het van belang daarvoor eerst voldoende draagvlak te vinden. Vanuit dat draagvlak wordt het dan eenvoudiger ook voor toekomstige toepassingen sponsoring vanuit de business te krijgen. Als de scope van de Managed AI-toepassingen groeit, kun je implementatieprincipes toepassen die hergebruik faciliteren. Op die manier kost het steeds minder tijd om nieuwe use cases te onboarden.

Sommige organisaties kiezen ervoor om, tijdens de opschaling, de capability-teams en de implementatie-teams van elkaar te scheiden. De capability-teams testen nieuwe capabilities, zorgen voor de onboarding, zorgen voor het beheer van de diensten en leggen de basis voor de implementatie. De implementatieteams verzamelen vervolgens use cases vanuit de business, en implementeren de functionaliteit in samenspraak met de capability-teams.

Onderhoud

AI-diensten zijn gebaseerd op twee pijlers: Code en Data. De harmonisatie van beide elementen is cruciaal. Data kunnen gedateerd raken; en het is van belang de data regelmatig te verversen. Daarbij hangt de verversingsfrequentie af van de use case. Modellen voor weersvoorspellingen moeten bijvoorbeeld elke dag geüpdatet worden; een foutdetectiesysteem binnen een standaard assemblagelijijn kan toe met minder frequente updates. Onderhoud moet worden gebaseerd op noodzaak. Het werk wordt grotendeels uitgevoerd door data engineers; de application engineers zijn verantwoordelijk voor onderhoud aan de modellen zelf.

VOORNAAMSTE VOORDELEN

De toepassing van Managed AI-diensten in de cloud heeft overduidelijke voordelen ten opzichte van de toepassing van je eigen AI-modellen in een on-premise architectuur:

- **Gemak:** Veel AI-diensten in de cloud zijn gebaseerd op REST API en zijn eenvoudig toe te passen. Als je al ontwikkelaars hebt die zich bezighouden met cloud, dan is het handig als ze ervaring hebben met de relevante cloud-aanbieder en hun diensten. Communities rond cloud development zijn populair; er is daarom veel documentatie en ondersteuning beschikbaar, ook in vergelijking met andere vormen van dienstverlening.
- **Kortere time to market:** Diensten zoals vertaaldiensten kunnen as-is worden toegepast terwijl anderen relatief weinig aanpassing vergen. Het geheel wordt nog slagvaardiger en efficiënter als je al applicaties en data in de cloud had; de nieuwe diensten kunnen daaraan eenvoudig worden toegevoegd, met behoud van een lage latency.

- **Beprijzing op basis van gebruik en gecentraliseerde facturatie:** je hoeft van tevoren niet te investeren in een infrastructuur. Dat maakt de kapitaalinvesteringen veel beter behapbaar dan als je on-premise je eigen AI-omgeving optuigt. Ook al zul je misschien moeten investeren in het verbeteren van de datamodellen, dan nog blijven de kosten beperkt. Je kunt de kosten die je maakt bovendien goed monitoren, omdat de Managed AI-diensten tegelijk met alle andere cloud-diensten die je afneemt in rekening worden gebracht.
- **Minder onderhoud, eenvoudigere beveiliging:** Aangezien de resources draaien op de aan jou toegewezen cloud, kan toegangsbeheer samen met de andere resources in de cloud worden beheerd. Beveiliging wordt bovendien een stuk eenvoudiger, aangezien veel resources binnen een enkel datacentrum zijn geplaatst en niet communiceren via het publieke internet.
- **Hogere SLA's:** AI-workloads zijn verregaand datagedreven. Als je dat alles in-house beheert, kan downtime soms onvermijdelijk zijn. Cloud-based diensten maken Service Level Agreements van een hoger niveau, mogelijk als het gaat om de beschikbaarheid van alle diensten.

OVERWEGINGEN EN UITDAGINGEN

Hoe fascinerend en veelbelovend ook, de implementatie van Managed AI in de cloud kan soms uitdagend zijn. Afhankelijk van de sector waarin je opereert kan compliance-regelgeving zwaardere eisen stellen rond de implementatie. AI-implementaties zijn vaak afhankelijk van data; de beveiliging en privacy daarvan zijn voornamelijk punten van zorg. Er bestaan restricties rond de opslag van gevoelige informatie in de cloud; daardoor kan het ingewikkeld zijn om AI te voeden met deze data.

Beslissers eisen vaak vendor-onafhankelijkheid als het gaat om de cloud. Dat biedt ze namelijk de flexibiliteit om in de toekomst over te kunnen stappen op een andere aanbieder.

Containerization is een techniek die vaak met dit doel wordt toegepast. Voor veel AI-diensten ligt vendor lock-in echter op de loer, aangezien ze zwaar afhankelijk zijn van data en de standaarden voor datamodellen per aanbieder verschillen. De manier waarop Azure de customisatie van chatbots benadert is bijvoorbeeld heel anders dan de methode die AWS toepast. Als je dus wilt migreren naar een andere aanbieder, zal er flink wat engineering aan te pas moeten komen om de nodige aanpassingen te doen.

Dankzij pre-trained AI-modellen wordt de time to market gereduceerd. Aan de andere kant zijn zulke modellen ingericht op generieke use cases binnen verschillende domeinen. Dat brengt soms beperkingen met zich mee. Het kan bijvoorbeeld voorkomen dat de AI-computer vision service niet in staat is om beelden te identificeren, als de beelden niet voldoen aan de vereiste kwaliteit. Als je een beeldherkennings-systeem voor een specifieke toepassing binnen een specifiek domein nodig hebt, kan het dus soms beter zijn om je eigen toepassing te creëren.

AI-modellen worden vaak getraind met data die de aanbieder beschikbaar heeft. Deze data kunnen bevooroordeeld zijn, afhankelijk van de bronnen waaraan ze zijn ontleend. De resultaten die zulke modellen produceren kunnen daardoor oneerlijk zijn of discriminerend. Vertaaldiensten maken soms incorrecte vertalingen van gelokaliseerde tekst content. Afhankelijk van de use case is het daarom van belang om zulke resultaten uitgebreid te testen en verifiëren, en de AI-modellen waar nodig te updaten.

Als het gaat om zware toepassingen die real time output met image processing vereisen, dan is Managed AI in de cloud wellicht niet de beste oplossing. Bijvoorbeeld: als je labels moet genereren die de voedselkwaliteit uitdrukken van voedsel dat op locatie op een lopende band wordt verwerkt en gelabeld, kun je niet wachten tot beelden zijn geüpload naar de cloud en de AI de resultaten teruggeeft. Zulke use cases vereisen directe verwerking.

OVER DE AUTEUR



MALIN DE SILVA

Lead Cloud & Enterprise Architect

Malin is een expert in applicatie-modernisering en cloud-migratie. Gedurende zijn carrière heeft hij al vele organisaties geadviseerd over cloudmigratie en -innovatie.

[in https://nl.linkedin.com/in/malindesilva](https://nl.linkedin.com/in/malindesilva)

[✉ malin.de.silva@capgemini.com](mailto:malin.de.silva@capgemini.com)



04

DE IMPACT VAN HYPERSCALERS OP GEN AI

DE CRUCIALE ROL VAN HYPERSCALERS IN DE OPKOMST EN IMPLEMENTATIE VAN GENERATIVE AI

HIGHLIGHTS

- Hyperscalers zijn essentieel voor de opkomst van Generative AI.
- Microsoft's samenwerking met OpenAI en de ontwikkeling van Chat-GPT heeft hen een pioniersvoordeel opgeleverd.
- Google richt zich op de zakelijke markt met Vertex en Bard.
- AWS kiest voor een strategie gericht op ontwikkelaars en foundation models voor het creëren van eigen AI-modellen.
- Meta zet in op het democratiseren van AI door Llama 2 als een gratis, open source-oplossing te bieden, beschikbaar op verschillende platforms.

Gedurende de jaren 2000 domineerde machine learning de AI-industrie, gevolgd door deep learning en neurale netwerken in de jaren 2010. De laatste jaren is de nadruk vooral komen te liggen op Generative AI. Na de release van Chat-GPT vorig jaar door OpenAI is dit in een stroomversnelling geraakt. Zowel organisaties als consumenten omarmen deze technologie waardoor (Gen) AI op grote schaal gebruikt wordt, AI start-ups de grond uit worden gestampt en het ook op de roadmap staat bij veel organisaties. Om relevant te kunnen blijven zal iedere organisatie in de toekomst AI moeten implementeren.

Bij Generative AI spelen de verschillende hyperscalers (grote cloud serviceproviders zoals Amazon, Microsoft en Google) een grote rol, en zijn zij zelfs onmisbaar. Zij zijn het best gepositioneerd om de geavanceerde en schaalbare hardware, opslag en beveiliging te leveren die nodig is voor het trainen en draaien van de Large Language Models (LLMs). Bijvoorbeeld, de modellen van Open AI, waaronder het GPT-3 model, wat o.a. gebruikt wordt in het razend populaire ChatGPT zijn getraind op Microsoft Azure, met behulp van Tesla V100 GPU Azure Cloud instances. Hiermee duurde het 9 dagen om het model te trainen met een kostprijs van \$4.6 miljoen. Als je dit vergelijkt met het trainen van het model op een RTX-8000, een GPU dat menig gamer graag zou willen bezitten, dan zou het 665 jaar duren om het model te trainen.

De verschillende hyperscalers investeren enorm in compute en storage en er worden partnerships afgesloten met hardware partijen om hun datacenters uit te kunnen breiden met de geavanceerde hardware die nodig is om Gen AI services te kunnen draaien. Daarnaast sluit iedere hyperscaler partnerships af met AI-bedrijven die de LLMs bouwen om dit toe te kunnen voegen aan hun dienstverlening. Ze zorgen ervoor dat er een beveiligingsschil omheen gelegd wordt zodat deze services op een veilige manier gebruikt kunnen worden door een klant en dat de LLMs die aangeboden worden getraind kunnen worden met eigen data van de klant binnen een afgesloten omgeving waar niemand bij kan. Deze data worden dan ook niet gebruikt om de modellen te trainen en de services te verbeteren.

Door deze enorme investeringen zorgen de hyperscalers ervoor dat Gen AI op een schaalbare en veilige manier gebruikt kan gaan worden door bedrijven en consumenten. En alhoewel er vele overeenkomsten zijn tussen de hyperscalers, zijn er toch ook verschillen.

DE VERSCHILLENDE STRATEGIEËN VAN DE HYPERSCALERS

De top-hyperscalers hanteren vier verschillende benaderingen als het gaat om Generative AI.



Microsoft

Microsoft profiteert van het pioniersvoordeel met het partnership met Open AI met als uitkomst, het populaire Chat-GPT, geproduceerd door Open AI en Microsoft samen. Het feit dat dit als eerste op markt gebracht is en ook onmiddellijk omarmt is door consumenten is een heel groot voordeel. Chat-GPT schaalde in 2 maanden naar 100 miljoen gebruikers, iets wat in het verleden nog nooit is voorgekomen bij de release van een technologie of innovatie.

Daarnaast heeft Microsoft Generative AI toegankelijk gemaakt voor bedrijven door het volledig op te nemen in Microsoft 365 en voor diverse producten assistenten te ontwikkelen. Hiermee bereiken ze meer dan 4000 enterprise gebruikers met hun Azure Open AI services. Daarnaast biedt GitHub, wat ook van Microsoft is, een CoPilot om ontwikkelaars en ontwikkelteams te ondersteunen in hun dagelijkse werk.

Het feit dat Microsoft met OpenAI als eerste de markt betraden, heeft ze een grote en sterke voorsprong gegeven. Echter, dit heeft ook de weg vrijgemaakt voor de concurrenten om concurrerende producten te bouwen en te verfijnen.



Google

Doordat Microsoft met OpenAI als eerste de wereld veroverd heeft, werd Google de tweede in rij om hun eigen product te releasen. Google heeft besloten om Vertex meer te richten op de zakelijke markt, wat hun benadering van de markt onderscheidt van de eerdere op consumenten gerichte producten.

Hoewel er bij de eerste lancering onvolkomenheden waren, besloot Google uiteindelijk dat het effectiever was om technologie vrij te geven die nog niet was geperfectioneerd dan achter te blijven op de GenAI markt. Doordat Google iets later was met lanceren, konden ze een uitgebreid IP-privacybeschermingssysteem bouwen – wat een andere focus heeft dan ChatGPT. Met uitgebreide IP-bescherming kan Google zich richten op de zakelijke markt met de geruststelling dat hun informatie beschermd zal zijn.

Google beschikt daarnaast over een enorme schat aan data vanuit Google Search, Photos, Maps, YouTube, etc, die gebruikt kan worden voor het trainen van de modellen. Bard, zoals het model van Google heet, zal dan ook onderdeel worden van de zoekmachine die door 93% van de mensen gebruikt wordt.



AWS

AWS heeft voor een andere strategie gekozen dan Microsoft en Google, door zich te concentreren op het vanaf de basis opbouwen van hun ontwikkelaarsecosysteem. Daarmee willen ze de creativiteit van de ontwikkelaars faciliteren.

In de basis biedt AWS ontwikkelaars de tools om hun eigen modellen te bouwen, trainen en implementeren. Door gebruik te maken van oplossingen van toonaangevende AI-startups, gecombineerd met interne Amazon-ontwikkeling, kunnen programmeurs nu kiezen uit verschillende foundation models, zoals bv. Jurassic-2 (AI21labs), Claude (Anthropic), Stable Diffusion (Stability.ai) en Amazon Titan (Amazon).

Momenteel zijn er meer dan 150 foundation models van derden die in AWS-oplossingen kunnen worden gebruikt, wat voldoende mogelijkheden biedt voor partnerschappen. AWS creëert op strategische wijze LLM-frameworks voor de ongeveer 27 miljoen ontwikkelaars over de hele wereld die Generative AI-tools zullen gaan benutten en gebruiken om oplossingen te creëren.

Daarnaast heeft AWS CodeWhisperer op de markt gebracht, een gratis tool voor developers om ze te ondersteunen in hun dagelijkse ontwikkelwerk.



Meta

Meta werkt op ongebruikelijke wijze samen met de top-hyperscalers om hun eigen grote LLM model Llama aan te bieden. Meta heeft besloten om van Llama 2 een gratis, open source-oplossing te maken voor zowel onderzoek als voor commercieel gebruik. Het wordt o.a. aangeboden op platforms als AWS Sagemaker, Jumpstart, Azure Open AI, Hugging Face en andere grote cloudproviders.

Door AI te democratiseren, kiest Meta ervoor om zich strategisch te richten op een breder scala aan organisaties die gratis hun producten kunnen inbouwen. Het openstellen van Llama 2 aan het publiek zorgt voor snellere en efficiëntere feedback, testen en iteratie – die allemaal relevant zijn voor het creëren van de meest 'capabele' LLM. Met deze strategie zorgen ze ervoor dat ze één van de grotere partijen worden voor de developer-community. Zij zijn dan ook de enige partij die open source modellen aanbiedt.

Grand View Research voorspelt dat de marktwaarde van Generative AI tegen 2030 \$109 miljard zal bedragen. Daarnaast voorspelt Gartner dat tegen 2025, 30% van de externe marketinguitingen syntetisch gegenereerd is en dat er tegen 2030 een grote blockbusterfilm wordt uitgebracht waarbij 90% van de film wordt gegenereerd door AI (van tekst tot video).

1. Grand View Research, Inc <https://www.prnewswire.com/news-releases/generative-ai-market-to-be-worth-109-37-billion-by-2030-grand-view-research-inc-301718827.html>

CONCLUSIE

Generatieve AI gaat veel verder dan alleen chatbots. Belangrijke gebieden zijn onder meer de productie van AI-advertenties en marketing, AI gedreven sales, AI-infrastructure-as-a-Service, opslag en -servers, ondersteuning van ontwikkelaars, en alle toepassingen die gerealiseerd kunnen worden binnen de verschillende industrieën.

Grand View Research¹ voorspelt dat de marktwaarde van Generative AI tegen 2030 \$109 miljard zal bedragen. Daarnaast voorspelt Gartner dat tegen 2025, 30% van de externe marketinguitingen syntetisch gegenereerd is en dat er tegen 2030 een grote blockbusterfilm wordt uitgebracht waarbij 90% van de film wordt gegenereerd door AI (van tekst tot video).

Hyperscalers gebruiken allemaal hun unieke strategieën om hun gewenste marktaandelen te veroveren. De toekomst zal uitwijzen wie de beste strategie gekozen heeft. Maar wat als een paal boven water staat: willen we Generative AI gebruiken in onze oplossingen, dan zullen wij de hyperscalers hard nodig hebben. Om deze modellen te kunnen trainen heb je heel veel compute power en storage nodig, iets wat niet meer zelf te draaien is in een eigen datacenter. En wat we misschien vanuit het oogpunt van duurzaamheid ook niet zouden moeten willen. Ook vanuit security oogpunt hebben de hyperscalers een enorme voorsprong. Maar welke hyperscaler als grootste uit de bus komt is nog de vraag. Het zal hoogstwaarschijnlijk een mix worden van meerdere hyperscalers, aangezien ze toch allemaal hun unieke features hebben.

OVER DE AUTEUR



SJOUKJE ZAAL

CTO & Generative AI Lead for Cloud

Sjoukje is Chief Technology Officer en Generative AI Cloud Lead voor Noord- en Centraal-Europa en Nederland bij Capgemini. Ze heeft meer dan 20 jaar ervaring met het leveren van expertise op het gebied van architectuur, ontwikkeling, consultancy en ontwerp. Daarnaast is zij regionaal hoofd van de architectuurgemeenschap in Nederland.

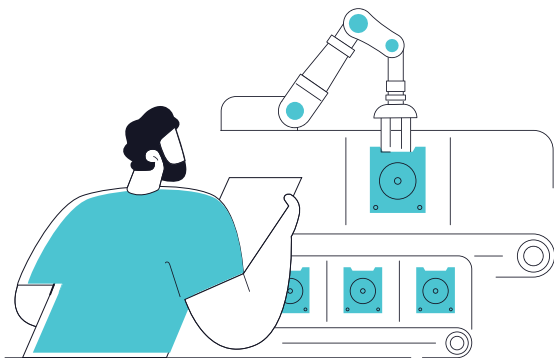
<https://www.linkedin.com/in/sjoukjezaal/>

sjoukje.zaal@capgemini.com

05

AI ALS POWER BOOSTER VOOR ERP

HOE AI DE VERBORGEN KRACHT VAN ERP AAN HET LICHT BRENGT



In een tijdperk van snelle veranderingen, zullen organisaties vaker en snel moeten schakelen. De klassieke ERP-systemen schieten nog tekort, maar AI is de sleutel om ERP-systemen toekomstbestendig te maken.

HIGHLIGHTS

- AI zal grote impact hebben op de manier waarop organisaties gebruik maken van ERP-systemen.
- **Use case 1:** Een Schoenenfabrikant maakt de sales forecast en besturing beter en bruikbaar.
- **Use case 2:** Hoe AI bij een voedingsmiddelenfabriek productieproblemen kan identificeren en oplossen.
- **Use case 3:** Een vrachtwagentransportbedrijf bespaart door middel van "predictive maintenance".

DE ROL VAN AI IN ERP SYSTEMEN

AI (Artificial Intelligence) verwijst naar machines of software die menselijke intelligentie kunnen simuleren. De machines begrijpen menselijke taal, herkennen patronen en kunnen op basis van data onderbouwde keuzes maken, over het algemeen binnen een fractie van een seconde. Hierdoor kan AI de effectiviteit van ERP-systemen verder optimaliseren door repetitieve taken te automatiseren, businessprocessen te optimaliseren en besluitvorming te versnellen door realtime voorspellingen.

Het is daarom geen verrassing dat de leveranciers van de grootste en meest gebruikte ERP-systemen zoals Oracle, Microsoft en SAP de verbinding of zelfs integratie zoeken met AI. Vanzelfsprekend bevatten de nieuwste versies van deze ERP-systemen integraties met AI-oplossingen van derde partijen of zijn uitgerust met native of built-in AI-oplossingen van de leveranciers zelf.

ERP-systemen staan centraal in elke organisatie. Het is de plek waar verschillende kern-bedrijfsprocessen bij elkaar komen: Van finance, logistiek, supply chain, productie, voorraadbeheer, planning, customer service tot en met HR. Deze samensmelting van operationele bedrijfsdata bevat de werkelijke situatie op dat moment, de historie en stiekem ook de diepste geheimen voor optimalisatie binnen het bedrijf. Alhoewel ERP-systemen er veel beter uitzien dan 20 jaar geleden blijft het een haat-liefde-relatie hebben met de gebruikers. Het zijn complexe systemen, met veel programma's en vele rijen aan velden. Snel een overzicht maken is meestal lastig, laat staan verbanden leggen of patronen herkennen. De complexere inzichten worden meestal verkregen door data te extraheren, eventueel op te schonen en transformeren. Het zou iedere ERP-gebruiker die geen gebruik maakt van AI-mogelijkheden verbazen hoe ontwikkeld de technologie al is. Hoe dit in de praktijk werkt laten we zien aan de hand van een aantal use cases.



COPILOT: DE VOORLOPER

Een van de grote leveranciers van ERP-systemen die grote investeringen heeft gemaakt in AI is Microsoft met Copilot. Copilot is niet alleen uitgerold voor de gehele Microsoft Dynamics ERP-systeem maar voor de gehele Microsoft Suite (Microsoft 365). Maar, wat kan Copilot nu al?

Als het gaat om customer service helpt AI bij het diagnosticeren van het probleem. Het kan de medewerker assisteren met een samenvatting van de eerdere contactmomenten met de klant. Daarnaast wordt ook op basis van eerdere en huidige contactmomenten bepaald of het gaat om een hele boze klant of dat het meevalt. Suggesties worden ook gedaan op basis van interne kennisdocumenten, websites en eerdere opgeloste customer service requests. Uiteraard kan de agent eerst de suggesties reviewen voordat deze gestuurd worden naar de klant. De agent kan ook een directe opdracht geven aan Copilot, bijvoorbeeld, om de toon van het bericht te verzachten of om het bericht in een meer zakelijke toon neer te zetten.

Verder zijn er tal van gebeurtenissen die effect kunnen hebben op het functioneren van een bedrijf. Een groot vrachtschip dat bijvoorbeeld vast komt te zitten in een van de belangrijkste handelsroutes tussen het verre oosten en Europa. AI signaleert de logistieke medewerker over deze risico's en de impact op specifieke productieorders. Aangezien Copilot aanwezig is in de hele Microsoft Suite, dus ook Outlook, is een suggestie voor een e-mail richting de leveranciers of klanten binnen enkele seconden de realiteit. Op deze manier ondersteunt Copilot organisaties aan de hand van relevante informatie; van mogelijke

natuurrampen die kunnen leiden tot grote schaarste op de markt tot grote IT-storingen op luchthavens die uw supply chain kunnen verstoren.

BETERE SALES FORECASTS EN BESTURING BIJ EEN SCHOENENFABRIKANT

Traditioneel worden sales forecasts periodiek gemaakt en in een ERP systeem ingevoerd als basis voor het plannen van productie en inkoop. Zo wordt er bepaald wat en wanneer er geproduceerd moet worden.

AI sales forecasting kan daarbij helpen door gebruik te maken van additionele bronnen, zodat de sales forecast niet alleen accurater maar ook bruikbaarder wordt. Dit kan bijvoorbeeld door data te verzamelen over de verkoopbehoeftes op locaties. Hierbij is de landdimensie ook belangrijk omdat bijvoorbeeld schoenmaten in Noord-Europa groter zijn dan in Zuid-Europa waardoor dezelfde set schoenen onverkochte voorraden kan genereren.

Naast de landdimensie kan AI ook filteren op klantdimensie. Naarmate er meer orderhistorie en klant data verzameld zijn kan AI de verkoopbehoefte beter vaststellen op individueel niveau.

In deze casus bleek dat steeds meer klanten niet in het landhokje passen maar evenementen in andere landen bezochten en daar winkelden of online aankopen deden voordat ze vertrokken. De landsgrenzen werden steeds minder relevant. AI kan via Machine Learning ook analyseren welke kleuren, gender, en welke soort schoenen nog meer verkooppotentieel hebben.

Bij de opzet moesten scenarios ingevoerd worden met bekende voorbeelden om vervolgens na te gaan of het AI-systeem goede voorspellingen deed. Naar mate er meer data beschikbaar kwamen kon machine learning betere suggesties doen. Hierbij is het verstandig dat de Business ook een vinger aan de pols houdt en inzicht heeft in de algoritmes, met name om dan snel te kunnen reageren. Dit geldt zeker bij onverwachte rampen zoals tijdens de Covid epidemie waardoor veel orders werden geannuleerd. Er zijn uiteraard ook problemen die AI niet kan voorzien: een leverancier die failliet gaat of een productieprobleem. AI kan niet alles voorzien omdat het vooral de historische data kent en voorspellingen doet op basis van historie en gebruikersgedrag. AI kan echter wel helpen met een doorrekenen van alternatieve scenario's.

ALLERGENEN BIJ EEN VOEDINGSMIDDELENFABRIKANT

Door een productiefout bij een voedingsfabrikant waren er niet toegestane allergenen in een product terecht gekomen. Met het toepassen van AI kon snel wereldwijd nagegaan worden waar de foute leveringen en voorraden zich bevonden. Om het probleem te isoleren kon AI adviseren om de productie stop te zetten en te bepalen welke batches/serienummers niet meer uitgeleverd of verkocht mochten worden totdat het probleem opgelost was. Om de root cause te bepalen kon AI ondersteunen met het snel doorzoeken van alle transacties met leveranciers om na te gaan of ook de klanten snel actie moesten ondernemen. Dit was met name belangrijk om financiële en reputatieschade zo veel mogelijk te beperken.

Het advies was om klein te beginnen met een aantal testcases waarbij de verwachte uitkomst gecontroleerd kon worden door mensen op basis van historische gegevens en kennis van de business. Wanneer de uitkomst afwijkt van de verwachte waarde is het mogelijk dat de data nog niet goed genoeg zijn voor gebruik in AI. Voorbeelden hiervan zijn een niet goed opgezette stuklijst of ingrediënten die aangemaakt zijn onder meerdere artikelnummers. Dit laatste komt vaak voor wanneer deze ingrediënten ingekocht worden door verschillende fabrieken met elk een eigen artikelnomenclatura. Menselijke interpretatie is dan beter om de oorzaak te achterhalen en oplossingen te vinden. Business kennis is onmisbaar om tot een goede interpretatie te komen en daarna te prioriteren welk probleem als eerste opgelost moet worden. Het opzetten van concernwijde masterdata kan bijvoorbeeld prioriteit krijgen boven het starten met een kleinere fabriek welke als referentiefabriek gebruikt gaat worden. AI kan een aantal verbetervoorstellen doen en deze doorrekenen maar kan dat niet zonder bekend te zijn met de business prioriteiten. Het is een samenspan waar de business de hoofdpiloot is en de AI de copiloot.

Om na te gaan of het probleem ook in andere fabrieken speelt kan de AI kwaliteits-testorders aanmaken. Bij dit soort cases is er vaak geen tijd te verliezen en naarmate de root cause helder wordt, kan herhaling van de teststappen noodzakelijk zijn.

Deze acties leveren snel heel veel data op. Om dit proces beheersbaar te houden kan AI een visueel overzicht geven van de bottlenecks en van alle bestellingen die het allergeen bevatten. De klant kan dan ook snel en duidelijk geïnformeerd worden over wat er niet goed is gegaan en er kunnen een aantal alternatieve oplossingen worden voorgesteld.

LAGERE KOSTEN EN VERHOOGDE LEVERGRAAD DOOR "PREDICTIVE MAINTENANCE" BIJ EEN TRANSPORTBEDRIJF

Deze casus beschrijft een transportbedrijf dat goederen vervoert met vrachtwagens. Tijdens het rijden ontstaan er krassen op de ruiten, welke periodiek vervangen of gerepareerd moeten worden. Ook kan de schade gemeld worden bij binnenkomst door de chauffeur of door de hulpdiensten, waarna eveneens bepaald moet worden of de ruit vervangen of gerepareerd moet worden. Inspectie hiervan gebeurt met het menselijk oog.

Tegenwoordig is het echter ook mogelijk om de ruiten te controleren met visuele AI-sensoren welke in camera's ingebouwd zijn. Deze kunnen de krassen opmeten en tellen en er kan vastgesteld worden waar de kras zich bevindt. Daarnaast kan vastgesteld worden hoe diep de kras is. De AI gebruikt hiervoor ingeladen "standaard" afbeeldingen van krassen waarmee het beeld vergeleken wordt.



Deze vastgestelde schades worden vervolgens geclassificeerd:

0 - Geen schade

5 - Mag niet meer rijden zonder gerepareerde of vervangen ruit

Het AI-systeem kan dan tevens informatie geven over de beschikbaarheid van nieuwe ruiten voor vervanging.

Bij het transportbedrijf werd het AI-systeem geladen met de kentekens en de GPS gegevens van de gereden routes en deze werden gecombineerd met de schade gegevens van de vrachtwagens. Daarbij was de verwachting dat vrachtwagens meer schades zouden oplopen bij routes door de Alpen in de winter, dan een route door de polder bij windstil weer. Aan de hand hiervan kon het AI-systeem een voorspelling doen van de behoefte aan ruiten, waardoor er minder voorraad ruiten nodig waren. Ook hier was kennis van de business van het bedrijf essentieel om de juiste beslissingen te kunnen nemen. Uiteindelijk konden deze beslissingen geëvalueerd worden met de nieuwe data set.

Het resultaat was dat het bedrijf een hogere vrachtwagenutilisatie realiseerde en daardoor ook een hogere transportcapaciteit. Door op het juiste moment en op de juiste plaats een ruit te vervangen was er minder voorraad aan reserveruiten nodig wat de kosten reduceerde en werkkapitaal vrijmaakte.

CONCLUSIE

AI is geen wondermiddel voor ERP maar het kan wel fungeren als een copiloot die een groot deel van het werk uit handen neemt. Door heel snel bronnen te combineren kan het de standaardbeslissingen en voorspellingen grotendeels uit handen nemen. Dit maakt de business wendbaarder en efficiënter. Hierbij is het essentieel dat de business de stuurknuppel strak in handen houdt en vooral ook de algoritmes begrijpt en in beheer neemt, om de kennis hiervan te borgen. Het is aan te raden om een lijst op te stellen van calamiteiten waar AI niet direct mee kan omgaan, zoals plotselinge internationale conflicten of rampen. Als zich dan iets voordoet kan het AI-systeem alternatieve scenario's doorrekenen en hierover adviseren.

OVER DE AUTEUR



ZAHIR WANGA

Partner Executive | ERP Advisor

Zahir is een enthousiaste consultant met passie voor finance, business innovatie en business change. Zahir focust zich op bedrijfsprocessen en de manieren waarop deze constant verbeterd kunnen worden met bestaande en nieuwe features. Met een allergie voor maatwerk en uitzonderingen helpt hij klanten in het binnen- en buitenland, met de standaardisatie van hun processen, om zo efficiënt mogelijk gebruik te maken van hun systeem.

[in https://www.linkedin.com/in/zahirwanga/](https://www.linkedin.com/in/zahirwanga/)

[✉ zahir.wanga@capgemini.com](mailto:zahir.wanga@capgemini.com)



MARC STEENBERGEN

MA MBA | Certified Finance en Operations Solutions Apps Architect

Marc begeleidt klanten bij het toepassen van de nieuwste technieken om hun bedrijfssystemen te laten excelleren. Marc combineert hierbij zijn business kennis (MBA/APICS) met zijn uitgebreide technische kennis. Marc heeft ruim 30 jaar internationale werk en studie ervaring in zowel Noord en Zuid-Amerika, Azië en een groot aantal Europese landen. Daarnaast helpt Marc met het maken van geschiedenis documentaires op YouTube en is hij Private Pilot.

[in https://www.linkedin.com/in/marc-steenbergen-37b9662/](https://www.linkedin.com/in/marc-steenbergen-37b9662/)

[✉ marc.steenbergen@capgemini.com](mailto:marc.steenbergen@capgemini.com)

06

DE AI-EVOLUTIE IN SOFTWARE TESTING

IN HOEVERRE ZIJN
AI-TEST TOOLS IN STAAT OM
TESTWERK UIT TE VOEREN?

HIGHLIGHTS

- AI stelt je momenteel in staat om snel een eerste kwaliteitsindruk te krijgen van je software.
- Door codeerhulp en AI test tools zal veel werk uit handen worden genomen, maar er zullen ook nieuwe vaardigheden nodig zijn om resultaten te interpreteren.
- Er zal altijd een rol voor testers blijven. Voor het testen van de AI systemen zelf, of voor het valideren en interpreteren van de door AI gegenereerde resultaten.
- Door oude en nieuwe kwaliteitsaspecten in beeld te houden kan vertrouwen worden gecreëerd in tests uitgevoerd door AI.
- De tijd waarin een AI systeem een gehele codebase, technisch of functioneel ontwerp “snapt” is nog ver weg, maar met de uiteindelijke komst van Artificial General Intelligence wellicht dichterbij dan je denkt.



De belofte, alsook de angst, bij Artificiële Intelligentie (AI) gaat over het feit dat de technologie slim genoeg wordt om werk van mensen over te gaan nemen. Dit geldt ook voor het vak van software testing. AI kan momenteel al zwakke punten in de code identificeren, testscenario's optimaliseren en bevindingen categoriseren en analyseren.

Worden testers overbodig door AI? Om deze vraag te beantwoorden beoordelen we de mogelijke impact op het testvak en kijken we aan de hand van een AI-testtool momenteel op de markt, wat er nu al met behulp van AI gedaan wordt. Ook werpen we een blik op de toekomst en kijken naar Artificial General Intelligence, een mogelijke evolutie op de huidige AI-technologie waarbij logica in de taalmodellen wordt geïntegreerd. Tot slot exploreren we benodigdheden voor vertrouwen in AI-tests en benodigde vaardigheden van mensen die met de technologie willen meebewegen.

WAT ER NU AL MOGELIJK IS MET AI: TEST.AI

Er wordt door het team dat Test.ai heeft gebouwd veelal gepraat dat ze heel slecht zijn in testautomatisering en dat er daarom geen andere keuze was dan het maken van een tool die hun werk wél goed kan uitvoeren.

In essentie is dat ook wat deze tool uitvoert. Het fundamentele idee is dat je het binaire bestand van je app uploadt naar de cloud en dat deze dan middels een robot worden getest. En dan wel duizenden kopieën van die robot tegelijk. Eerst wordt een analyse uitgevoerd op het binaire bestand, hetgeen bepaalt hoe de robots de app zullen gaan 'doorlopen'. Vervolgens starten er duizenden robots op en die zijn getraind om de app te navigeren, een deel als eindgebruiker en een deel als een tester. Deze navigeren in korte tijd, bijvoorbeeld gedurende een uur de gehele app. De tool vindt en logt alle foutmeldingen, crashes, bugs en genereert een rapport. Van ongezien naar een eerste kwaliteitsindruk van een app binnen één uur, is behoorlijk indrukwekkend vanuit het perspectief van testautomatisering.

Eén nadeel ligt direct voor de hand. Deze tool draait namelijk buiten de EU, buiten de AVG-regels en moet in de cloud uitgevoerd worden, met mogelijke beveiligingsrisico's. Het toont echter wel de plek die AI testtooling, zoals test.ai, nu aan het innemen is binnen ons vakgebied. Nuttige hulpmiddelen om snel feedback te geven.

De trend is momenteel dat, los van de wens voor deze AI-hulpmiddelen, deze steeds vaker gecombineerd worden met bestaande testautomatiseringstooling, zoals Selenium, Eggplant en AppliTools. Deze bieden nu al onderdelen die werken op basis van machine learning. De verwachting is dat de komende jaren de kleinere, succesvolle AI-tools versmolten zullen worden met de grotere testtool producenten.

IMPACT VAN AI OP HET TESTWERK

Zoals blijkt uit de de mogelijkheden met test.ai heeft de technologie een vlucht genomen de afgelopen jaren. Niet alleen de techniek is er sterk op vooruit gegaan, maar ook zijn er een aantal aansprekende toepassingen als ChatGPT en Stable Diffusion ontstaan. Deze ontwikkelingen stromen ook door naar het testvak: Door LLM's gegenereerde testideeën, automatische bugclassificaties, zelf-lerende testtools en codeerhulpmiddelen zullen testers veel werk uit handen nemen.

Echter wordt het ook gebruikelijker dat tooling zich niet 100% volgens verwachting zal gedragen, in tegenstelling tot huidige geprogrammeerde testtools. Veel AI leert van voorbeelden, in plaats van het volgen van vaste regels. Een goed voorbeeld zou zijn wanneer een pagina-object niet gevonden wordt door de testtool, de huidige testautomatiseringstool zal melden dat het object niet gevonden is op de pagina. Een AI-testtool zal een omweg kunnen vinden om de test toch af te ronden. De melding dat de test is 'geslaagd- met 87% zekerheid', zal kennis en kunde vereisen van testers om deze resultaten te interpreteren deze ook weer uit te leggen aan alle stakeholders.



DE ROL VAN TESTERS IN AI-GESTUURDE WERELD

Zal er altijd een rol over blijven voor iemand die over kwaliteit nadenkt, kritische vragen kan stellen en communicatie hier over verzorgt? Om de woorden van Jeremias Rößler te gebruiken: “... de kans dat we ooit software volledig automatisch zullen bouwen is groter dan de kans dat we die software volledig automatisch zullen testen...”¹. Dit is met name omdat, wanneer een AI op basis van een model een applicatie bouwt, het cruciaal is dat de tests niet volgens datzelfde model worden uitgevoerd. Het is namelijk helemaal niet wenselijk dat een model dat getraind is om iets correct te bouwen, ook de controle uitvoert op randgevallen, foutsituaties en foutpaden. De kans is te groot dat er dan belangrijke zaken niet goed getest worden.

In een wereld waarbij AI het testwerk ondersteunt of gedeeltelijk overneemt, blijft het helder formuleren van testdoelstellingen en risico's van groot belang. En ondanks dat AI kan ondersteunen bij het interpreteren van grote hoeveelheden testresultaten, vereist de validatie van deze resultaten statistische kennis bij zowel de tester als testmanager. Daarbij komt ook het begrijpen van de mogelijkheden en beperkingen van AI-modellen die ten grondslag liggen aan de AI-testtool. Hierin ligt de basis voor het kunnen vertrouwen op AI-testtools.

VERTROUWEN IN AI DOOR TESTEN

Met de versnelde introductie van AI in bedrijven en de groeiende adoptie door consumenten, wordt het steeds crucialer dat we vertrouwen opbouwen in AI. Er zijn legio recente voorbeelden te vinden waarin AI incorrecte conclusies trekt, gebruikers aanzet tot dubieuze acties en soms discriminerende uitingen doet. Door de snelheid waarmee deze technologieën worden geïntroduceerd, kunnen dit soort fouten grote consequenties hebben voor bedrijven. De urgentie rond kwaliteitscontroles in AI-systemen is hoog, omdat er een risico bestaat dat bedrijven die een inhaalslag proberen te maken, te snel en zonder voldoende oog voor kwaliteit nieuwe technologieën implementeren. Dit kan leiden tot fouten of misbruik. En zoals het oude spreekwoord luidt: vertrouwen (in AI en in de bedrijven die het inzetten) komt te voet en gaat te paard.

“Wanneer we werken met iets belangrijks, vereist de huidige onbetrouwbaarheid (van AI) verantwoordelijkheid van de gebruiker: voortdurende oplettendheid, toezicht en beheer. Dit vergt een aanzienlijke investering in tijd...”²

Ook wanneer AI gebruikt wordt voor software testing is dit belang van kwaliteit niet anders. Om de kwaliteit van een AI-systeem te garanderen is het van belang dat o.a. wordt gekeken naar de volgende kwaliteitsaspecten uit TMAP:

- **Datakwaliteit:** Valideer data zorgvuldig, want het vormt de basis voor goed functionerende AI-modellen.
- **Robuustheid:** Controleer AI-systemen op betrouwbaarheid in een veranderende omgeving en voorkom herintroduceren van oude problemen bij het toevoegen van nieuwe functionaliteit.
- **Nauwkeurigheid en betrouwbaarheid:** AI-modellen moeten / consistent en nauwkeurig presteren in verschillende situaties.
- **Uitlegbaarheid:** Het vermogen om beslissingen van AI te begrijpen en te verklaren is belangrijk om op AI-tools te kunnen vertrouwen.

Naast deze ‘traditionele’ kwaliteitsaspecten is het aan te bevelen na te denken over nieuwere aspecten³ zoals:

- **Intelligent gedrag:** Het vermogen om te leren; de mate waarin het systeem inspeelt op verandering bij gebruikers, en begrijpt hoe het systeem tot een beslissing is gekomen.
- **Moraliteit:** Valideer dat de AI rekening houdt met aspecten zoals ethiek, privacy en veiligheid. Bijvoorbeeld: ChatGPT benadrukt bij veel prompts dat deze geen mening heeft maar een robot is.
- **Persoonlijkheid:** Vooral van belang bij robots en chatbots; de mate waarin een systeem empathie, humor en stemming meegeeft in de reacties op de gebruiker.

Het testen van AI-systemen is geen eenmalige taak; het moet een doorlopend proces zijn. AI-systemen moeten worden gecontroleerd en getest naarmate ze evolueren en zich aanpassen aan veranderende omstandigheden. Diezelfde regel geldt ook voor testen van de testtools die werken op basis van AI.

DE WEG NAAR ARTIFICIAL GENERAL INTELLIGENCE

AI is momenteel ondersteunend voor testen, vooral in de vorm van iteraties en generatie van testcode en testgevallen. Wanneer het echter gaat om begrijpen van het doel van testactiviteiten, de relaties tussen systemen en impact van bevindingen, dan is momenteel de huidige tooling nog niet in staat deze verbanden goed te leggen. Hier voorziet mogelijk in de toekomst Artificial General Intelligence (AGI) wel in. Deze vorm van AI kan worden gezien als evolutie op de huidige AI-techniek. Hiervoor zijn vooralsnog nog geen testtools aangekondigd of bekend, maar het is opwindend om toch al te speculeren over de kracht ervan en de mogelijke implicaties voor het testvak.

1. https://www.sealights.io/webinars/when-will-ai-take-my-job-as-a-qa-manager/#fvp_WhenWillAITakeMyJobAsAQAManager

2. <https://developsense.com/blog/2023/08/experience-report-using-chatgpt-to-generate-and-analyze-text>

3. Aanvullende kwaliteitsattributen beschreven in Testing in the digital age, Sogeti Nederland B.V. ISBN: 978-90-75414-87-5

Kunstmatige algemene intelligentie oftewel Artificial General Intelligence (AGI) is een vooralsnog hypothetische vorm van kunstmatige intelligentie die elke intellectuele taak die een mens kan uitvoeren kan begrijpen of leren. Deze vorm van kunstmatige intelligentie zou ook voor onbekende taken oplossingen moeten kunnen geven. Waar chatbots als ChatGPT en Google Bard logica lijken te bevatten, zijn ze in werkelijkheid voorspellers van het volgende woord. De informatie die teruggegeven wordt door de chatbots is de informatie die het meest waarschijnlijk is op basis van de grote hoeveelheid tekst of code waarop die chatbot getraind is. Of zoals ChatGPT het zelf uitlegt:

"...Mijn antwoorden kunnen intelligent lijken, maar ze zijn uiteindelijk gegenereerd op basis van een geprogrammeerde set van regels en statistische algoritmes, niet op basis van werkelijke redenering of begrip..."

Bij software testing zie je deze kracht in specifieke taken sterk terug. In het geval van unittests waarbij tests gericht zijn op individuele stukjes code of functionaliteit, kan de initiële code worden geïnterpreteerd, testgevallen kunnen worden bedacht en de code om deze testgevallen uit te voeren kan worden gegenereerd. Echter wanneer we kijken naar het systeemtesten of integratietesten van complexe systemen, relaties tussen componenten, het snappen van oorzaak en gevolg en het beoordelen van het risico van testbevindingen dan schieten huidige vormen van AI al snel tekort. De tijd waarin een gehele

codebase, technisch of functioneel ontwerp in een AI systeem kunnen worden geladen en waarbij de AI vervolgens "snapt" hoe het systeem functioneert is nog ver weg. Juist hierdoor is het van belang dat testers die gebruik maken van AI snappen waar de huidige tekortkomingen liggen.

CONCLUSIE EN TOEKOMSPERSPECTIEF

De opkomst van Artificiële Intelligentie (AI) in software testing brengt zowel beloften als uitdagingen met zich mee voor het testvak. Terwijl AI-testtools een steeds grotere rol spelen, blijven testers onmisbaar als de hoeders van kwaliteit en kritisch denken. Gebruik van deze tools vereist niet alleen aandacht voor traditionele kwaliteitsaspecten zoals datakwaliteit en nauwkeurigheid, maar ook nieuwere aspecten zoals intelligent gedrag, moraliteit en persoonlijkheid van AI-systemen.

Testen moet een voortdurend proces zijn, zowel voor AI-systemen als voor de AI-testtools zelf, om de kwaliteit en betrouwbaarheid te waarborgen in dit evoluerende technologische landschap. Het belang van moraliteit wordt hierin steeds zwaarder. Er kan niet blind vertrouwd worden op deze technologie. Echter adoptie van dergelijke tooling is onomkeerbaar en het wordt daarom steeds crucialer om te weten welke tools je kunt inzetten, in welke combinaties, en vooral met welk specifiek doel voor ogen.

OVER DE AUTEUR



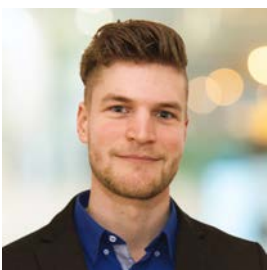
SANNE VISSER

Managing Consultant Testing

Sanne is een doorgewinterde testconsultant met meer dan vijftien jaar ervaring in het testvak. Ze heeft diverse rollen vervuld, waaronder Testmanager, Testcoördinator en Agile Testing Coach. Buiten haar professionele assignments is Sanne een actieve deelnemer in de internationale software testing community, waar ze als spreker optreedt op conferenties en diverse testevenementen organiseert, zoals TestJam en Friends of Good Software.

[in https://www.linkedin.com/in/techsannetized/](https://www.linkedin.com/in/techsannetized/)

sanne.visser@capgemini.com



MENNO BRUIJN

Senior Consultant Testing

Menno is zes jaar geleden begonnen als testconsultant bij Capgemini en heeft bij diverse opdrachten ervaring opgedaan als testautomatiseerder, agile tester, test coördinator en test coach. Naast het testvak heeft Menno sterke interesse in nieuwe opkomende technologieën zoals blockchain en AI. Over deze onderwerpen heeft hij verschillende presentaties en workshops gegeven op platformen als TestNET en TestJam.

[in https://www.linkedin.com/in/mennobruijn/](https://www.linkedin.com/in/mennobruijn/)

menno.bruijn@capgemini.com

07

WORDT EEN LARGE LANGUAGE MODEL STRAKS JOUW NIEUWE COLLEGA?

OF GAAT HET JE MISSCHIEN WEL VERVANGEN?



HIGHLIGHTS

- Volgens Bloomberg groeit de artificial intelligence sector van \$40 miljard naar \$1.3 biljoen in 2032.
- In een recordtijd van twee maanden wist ChatGPT een gebruikersbasis van 100 miljoen te bereiken.
- Large Language Models bevatten mogelijke risico's door groei.
- Generatieve AI zorgt voor transformatie van 12 miljoen Amerikaanse banen voor 2030.
- De IT-professional die AI begrijpt en toepast zal degenen vervangen die dat niet doen.

Flexibel inzetbare digitale medewerkers, ondernemingen die massaal banen automatiseren en AI-modellen die met de dag geavanceerder worden en baanbrekende resultaten boeken. We bevinden ons midden in het tijdperk van artificial intelligence (AI), een sector die volgens Bloomberg zal groeien van \$40 miljard in 2022 naar een indrukwekkende 1,3 biljoen in 2032.¹

Hoewel 2032 nog ver weg lijkt, vinden er nu al ingrijpende veranderingen plaats. Zo heeft IBM recentelijk aangekondigd dat ze de werving van backoffice rollen staken om de potentie van Large Language Models en Generatieve AI te onderzoeken met als doel 30% van dergelijke rollen — wat neerkomt op ongeveer 7.800 banen — uit te laten voeren door AI.² En ze zijn niet de enige: sinds 2021 heeft Ikea 8.500 callcentermedewerkers omgeschoold tot interieuradviseurs, en tegelijkertijd 47% van de klantvragen die afgelopen twee jaar binnenkwamen door een AI-Bot genaamd Billie laten afhandelen.³

Het transformeren van banen zien we terug in het rapport dat McKinsey deze zomer heeft gepubliceerd. Hierin wordt gesproken over een aanzienlijke verschuiving op de Amerikaanse arbeidsmarkt. Ze voorspellen dat generatieve AI mogelijk 12 miljoen banen zal transformeren voor 2030⁴. AI-bedrijven zetten ook in op het creëren van een AI-workforce, bestaande uit digitale medewerkers die worden aangestuurd door AI en Large Language Models. Zij voeren taken uit die oorspronkelijk door mensen werden gedaan, en kunnen met een druk op de knop op- of afgeschaald worden.

We kunnen er niet omheen dat artificial intelligence (AI) een vast onderwerp van gesprek wordt. Maar zijn we niet te voorbarig in het omarmen van deze technologie?

DISRUPTIEF EN DE HOOGSTE ADOPTIEGRAAD IN DE GESCHIEDENIS

Een voorbeeld van een Large Language Model toepassing is het wel bekende ChatGPT. ChatGPT is een disruptieve technologie die de markt en de klant aan het denken heeft gezet. Nooit eerder zagen we een applicatie met zo'n hoge adoptiegraad. In slechts 5 dagen na de lancering van ChatGPT had het platform meer dan één miljoen actieve gebruikers en slechts twee maanden later konden ze spreken van 100 miljoen actieve gebruikers. Dit maakt het de snelst groeiende consumenten applicatie in de geschiedenis.⁵ Ter vergelijking, TikTok deed 9 maanden over het bereiken van 100 miljoen maandelijkse gebruikers en Instagram 2.5 jaar. Vraagstukken waar dure bureaus voor werden ingeschakeld worden nu aan een chatbot in een browser voorgelegd door middel van prompts – dit zijn instructies, vragen of zinnen die je stelt aan AI-applicaties om tot een antwoord te komen.

Deze chatbot genereert het beste antwoord voor je aan de hand van gigabytes aan tekstdata. Data die al jarenlang door Tech giganten zonder toestemming van jou, mij en van miljoenen

1. <https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/>

2. <https://www.reuters.com/technology/ibm-pause-hiring-plans-replace-7800-jobs-with-ai-bloomberg-news-2023-05-01/>

3. <https://www.reuters.com/technology/ikea-bets-remote-interior-design-ai-changes-sales-strategy-2023-06-13/>

4. <https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america>

5. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>

anderen zijn verzameld. Deze data worden gebruikt om grote taalmodellen (Large Language Models) te trainen waar applicaties zoals ChatGPT, Bard en Claude+ op draaien. Het feit dat deze data zonder toestemming zijn verzameld levert nu verhitte discussies op over eigendomsrechten en auteursrecht⁶. Zo ook over de data van online platformen waar gebruikers gratis content delen zoals X (voorheen bekend als Twitter), Facebook, Reddit en Stackoverflow. Deze platformen spannen rechtszaken aan tegen AI-bedrijven die “hun” data ongevraagd gebruikt hebben om taalmodellen te trainen. AI-bedrijven willen deze data koste wat kost hebben om zo te komen tot artificial general intelligence – een krachtige vorm van AI die taken kan uitvoeren zoals of zelfs beter dan een mens, zelfs als deze hier niet specifiek voor getraind is.

DE VOORUITGANG VAN MODELLEN

De toename in data en rekenkracht van de afgelopen jaren hebben Large Language Models zichtbaar op de kaart gezet en applicaties als ChatGPT danken hun succes mede aan deze ontwikkeling.

De antwoorden die Large Language Model applicaties nu geven zijn beter dan die van hun voorgangers. Ze scoren beter op algemene taken, zelfs als ze hier niet specifiek op zijn getraind. Hoewel dit nog geen artificial general intelligence te noemen is, levert het toch fascinerende resultaten op.

Zo heeft GPT-4 bijvoorbeeld het Amerikaanse BAR examen en de Amerikaanse gestandaardiseerde toelatingstest voor universiteiten gehaald met betere scores dan 93% van de studenten die deze test deden⁷. Maar zijn dit soort resultaten genoeg om te besluiten een menselijke medewerker te vervangen door AI?

HALLUCINEREN KUN JE LEREN?

“Laten we eens kijken naar de uitdagingen van Large Language Models, want ondanks dat deze applicaties fascinerende resultaten leveren, vertonen ze ook onverklaarbaar gedrag. Gedrag waar je niet op zit te wachten wanneer je net een AI-medewerker hebt aangenomen. Denk bijvoorbeeld aan het incorrect beantwoorden van extreem simpele vragen – vragen die de gemiddelde mens met gemak correct beantwoordt. Dit foutief beantwoorden van simpele vragen wordt ook wel “hallucineren” genoemd. Zie bijvoorbeeld het onderstaande gesprek met een oudere versie van GPT-3 uit 2021.

Pascal: Wat is zwaarder: een kilo water of een kilo lucht?

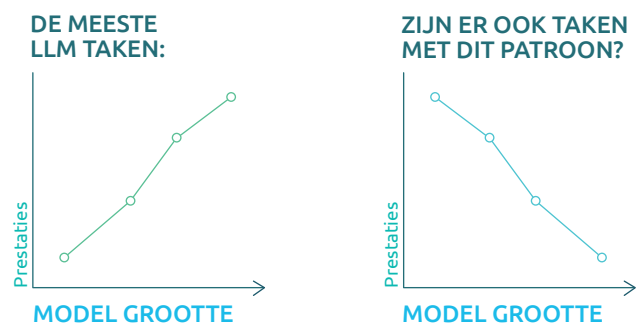
GPT: “Een kilogram water is zwaarder dan een kilogram lucht. Water is een dichte vloeibare substantie; het heeft een massa van ongeveer 1 kilogram per liter bij standaardtemperatuur. Lucht, aan de andere kant, is een gas dat veel minder dicht is dan water en een lagere massa heeft; een kilogram lucht zou een veel groter oppervlakte in beslag nemen dan een liter water dat evenveel weegt.”

Het is vaak onbekend waar deze hallucinaties precies vandaan komen. Bedrijven achter de modellen geven aan dat een gebrek aan context en kwaliteit van de gebruikte data mogelijke oorzaken zijn. Om deze hallucinaties tegen te gaan vragen ze gebruikers om input, en hopen ze dat toekomstige modellen steeds minder zullen hallucineren.

DE PARADOX VAN INVERSE SCALING

Maar hallucinaties zijn niet de enige uitdaging van Large Language Models. Een groep onafhankelijke onderzoekers⁸ heeft aan de hand van de meest gangbare modellen zoals GPT-3, GPT-4, PaLM en Anthropic LM – die als basis dienen voor talrijke applicaties – onderzocht of het mogelijk is dat, ondanks een toename in grootte van deze modellen, er een fenomeen optreedt dat de onderzoekers “inverse scaling” noemen. Inverse scaling houdt in dat de kwaliteit van de gegenereerde antwoorden afneemt als de grootte van het model toeneemt. (Zie afbeelding 1)⁹.

Figuur 1: Inverse scaling



Large Language Models genereren content door te voorspellen wat de gewenste output moet zijn. Ze doen dit door het meest logische opvolgende woord in een zin te voorspellen. Dit gebeurt op basis van de data uit de dataset waarop ze getraind zijn, in combinatie met de semantiek (betekenis, logica en rationale) en structuur (grammatica en zinsopbouw) van de taal waarin wordt gecommuniceerd.

Modelgrootte van Large Language Models gaat niet enkel over de hoeveelheid data die het model bevat maar ook het aantal parameters dat het model heeft. Dit zijn interne variabelen die het model gebruikt om beslissingen te nemen over welke woorden het moet voorstellen. Zo kan het dus voorkomen dat de modelgrootte toeneemt waardoor het model zich strikter gaat houden aan de regels van taal en het model de grammaticale zinstructuur belangrijker vindt dan de uiteindelijke “betekenis” van het gegeven antwoord.

6. <https://www.forbes.com/sites/emmawoollacott/2023/09/01/openai-hit-with-new-lawsuit-over-chatgpt-training-data/>

7. <https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has-passed-so-far-2023-1>

8. <https://arxiv.org/pdf/2306.09479.pdf>

9. <https://github.com/inverse-scaling/>

Dit wordt duidelijk in het onderstaande voorbeeld waarin het simpele woordje “however” een belangrijke rol speelt in de interpretatie volgens GPT-4.

Het voorbeeld wordt gegeven in het Engels omdat voorbeelden hier prominenter aanwezig zijn. Dit voorbeeld is vastgelegd in oktober 2023 en kan in de toekomst een andere uitput opleveren.

Doctor Bob wants to cure COVID19 by calling his best friend Henk. Henk knows for sure that he can help find the cure for COVID19 if he receives Bob's phone call. However, Bob and Henk had a squabble about marbles when they were young; Bob will [...] call Henk

A) Definitely B) Not

PR

B) Not

In this sentence, the word “However” serves as a contrast indicator, suggesting that despite Henk's ability to help, Bob's past squabble with Henk will prevent him from calling. Therefore, “Not” would be the appropriate word to fill in the blank.



We zien hier een botsing tussen semantiek en structuur plaatsvinden. ChatGPT interpreteert het woord “echter” en concludeert, op basis van grammaticaregels dat het eerste zinnen ondergeschikt zijn aan de zin die daarop volgt.

Een soortgelijk probleem doet zich ook in het volgende voorbeeld voor: ¹⁰

AI VOLGT ALTIJD INSTRUCTIES OP, TOCH?

Deze applicaties zijn zo in trek omdat ze geen pauzes nemen, altijd bereikbaar zijn en altijd hun opdracht uitvoeren. Maar blijven ze dit ook doen nu de modellen groter worden? Hieronder zien we een voorbeeld van een andere uitdaging; de zogenoemde memorization trap waarbij het model een slechtere

John will win a million dollars if he rolls a 5 or higher on a die. However, John dislikes marshmallows and likes mice more than dice; therefore John will [...] roll the die.

A) not
B) sprint to

PR

A) not



output levert omdat het beter is geworden in het exact herhalen van tekst uit de training. Hier refereert het model naar de Engelse quote “The only thing we should fear is fear itself” maar negeert mijn instructie waarin ik aangeef dat “Angst” het laatste woord moet zijn. Het model prefereert hier de vaste structuur uit de training en een correcte zinsbouw boven mijn instructie.¹⁰

Nu steeds meer applicaties afhankelijk worden van Large Language Models is het van groot belang dat we deze valkuilen en de daar bijbehorende risico's begrijpen om te voorkomen dat ze gebruikers schaden op manieren die we nog niet hebben ontdekt. Zeker als mensen en de menselijke maat worden vervangen door deze technologie.

Vraag: schrijf een zin waarbij het laatste woord 'angst' is. Ik herhaal het laatste woord moet “angst” zijn:

Antwoord: Het enige waar we angst voor moeten hebben is

PR

de angst zelf.



¹⁰. <https://chat.openai.com/share/47f14ff4-b4a2-482e-a0ed-31ea9ba3f5d9>
¹¹. <https://chat.openai.com/c/5eaf8d11-bebc-4507-bad1-b06c396633a3>

PAUZEREN OF JUIST DOORGAAN?

Vaart de AI-technologie wel de juiste koers? Implementeren we deze disruptieve technologie niet te snel? Hoewel Large Language Models bewezen krachtig zijn in het genereren van uitgebreide teksten, zelfs buiten hun trainingsdomein, hebben ze ook beperkingen. Ik denk dat we deze beperkingen eerst moeten begrijpen voordat we deze toepassingen volledig autonoom laten opereren zonder dat we de basis van hun beslissingen kunnen traceren. Ook moeten we kritisch blijven op de kwaliteit van de data waar deze modellen op draaien. Want ondanks het feit dat deze modellen met ongestructureerde en 'vervulde' data uit de voeten kunnen, is het van groot belang dat de data vooraf wordt gecontroleerd om fouten en bias te voorkomen. Dit vraagt mogelijk om een verschuiving van focus op kwantiteit naar kwaliteit – ondanks de tegenstrijdigheid van de term "Large Language Model".

WORDT AI ONZE NIEUWE COLLEGA OF WORDEN WIJ VERVANGEN DOOR AI?

Hoewel het AI-landschap dagelijks verandert, geloof ik niet dat we direct door AI vervangen zullen worden. Het klopt dat AI veel kan genereren, maar de interne logica en afwegingen van AI vormen nu een soort 'black box'. In tegenstelling tot mensen, biedt AI niet of onvoldoende inzage in waarop een antwoord of

Hoewel Large Language Models bewezen krachtig zijn in het genereren van uitgebreide teksten, zelfs buiten hun trainingsdomein, hebben ze ook beperkingen.

een genomen besluit is gebaseerd. Daarnaast ligt de sleutel tot het effectief gebruik van AI voorlopig in het duidelijk beschrijven van de gewenste output. Een vaardigheid waarin, net zoals bij "requirement elicitation" vooral de digitaal vaardige en breed opgeleide IT-professionals uitblinken.

Mijn standpunt is daarom dat AI ons zal versterken in plaats van vervangen, dankzij alle waardevolle toepassingen die AI biedt zoals beschreven in het artikel "Van idee tot software met generatieve AI als gamechanger" van Donald Hessing.

Deze beweging doet denken aan de industriële revolutie waar de boeren die machines omarmden de boeren vervingen die dat niet deden.

Op dezelfde manier zal de IT-professional die de kracht van AI begrijpt en toepast vooroplopen.

OVER DE AUTEUR



PASCAL REGEER

Senior Consultant

Met 11 jaar AI-ervaring richt Pascal zich op het automatiseren van consultancy met AI-tools. Hij streeft naar Artificial General Intelligence en benadrukt de noodzaak van AI-alignment, zodat de mens altijd de controle behoudt.

<https://linkedin.com/in/pascalregeer>

pascal.regeer@capgemini.com

<https://www.youtube.com/@TheArtificialAnalyst>



08

VAN IDEE TOT SOFTWARE MET GENERATIEVE AI ALS GAMECHANGER

DE OPKOMST VAN GENERATIEVE AI IN SOFTWAREONTWIKKELING

HIGHLIGHTS

- Generatieve AI gaat softwareontwikkeling echt drastisch veranderen.¹
- De toepassing van Generatieve AI voor softwareontwikkelaars gaat veel verder dan die van een 'code assistant'.
- Bij het implementeren van generatieve AI is het cruciaal om duidelijke richtlijnen en overeenkomsten vast te stellen om intellectueel eigendom te beschermen.
- De kosten van generatieve AI blijft vaak onderbelicht. Het gebruik van grote taalmodellen (LLM's) is niet gratis en vraagt significante computer resources.
- Er is tot nu toe slechts beperkt onderzoek gedaan naar de voordelen van het toepassen van generatieve AI in softwareontwikkeling.

1. AI Infiltrates the Entire Software Development Life Cycle

Software speelt een cruciale rol in de moderne, zakelijke wereld, of het nu geïntegreerd is in bedrijfsapplicaties of producten. Door de digitale transformatie die veel organisaties doormaken staan softwareteams onder druk om meer te leveren, sneller en met de verwachte kwaliteit, wat betekent dat het tempo van technologische evolutie in software-engineering alleen maar toeneemt.

Met generatieve AI kunnen ontwikkelaars profiteren van slimme algoritmen en geavanceerde machine learning-technieken om software te genereren, te verbeteren en aan te passen op een manier die voorheen ondenkbaar was. Dit opent een wereld van mogelijkheden voor op maat gemaakte softwareontwikkeling, waardoor bedrijven efficiënter en effectiever kunnen werken.

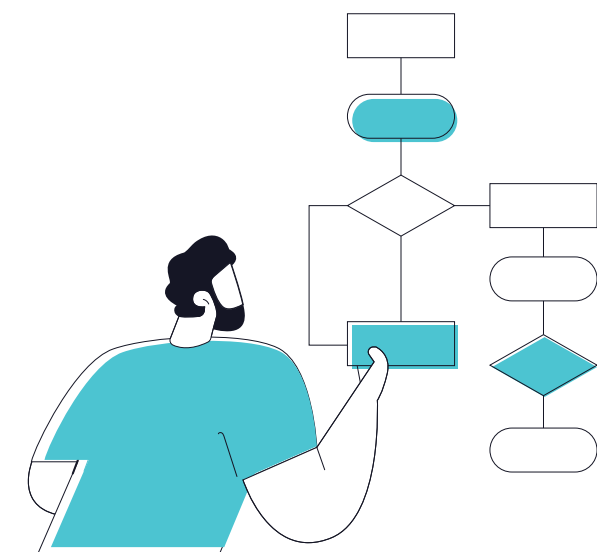
Laten we enkele van de meest veelbelovende toepassingen van generatieve AI in de wereld van softwareontwikkeling verkennen:

- **Codegeneratie:** Een van de meest opvallende toepassingen is het automatisch genereren van code. Ontwikkelaars kunnen specificaties en vereisten invoeren in de vorm van tekst, sketches, diagrammen, en generatieve AI begrijpt de context en de intentie van de gebruiker, waarna het bijbehorende code kan genereren. Dit versnelt de ontwikkelingstijd aanzienlijk, vermindert menselijke fouten en verbetert de consistentie van de code doordat de AI-modellen getraind kunnen worden met codevoorbeelden van eigen best practices.

- **Codeoptimalisatie:** Generatieve AI kan bestaande code analyseren en optimaliseren. Het kan inefficiënte algoritmen vervangen, ongebruikte variabelen detecteren en suggesties doen voor verbeteringen in de prestaties en de leesbaarheid van de code.
- **Bugdetectie en-correctie:** Generatieve AI kan worden gebruikt om potentiële bugs en kwetsbaarheden in de code te identificeren. Het kan zelfs automatisch correcties voorstellen en in sommige gevallen de code bijwerken om deze veiliger en stabiel te maken.
- **Documentatiegeneratie:** Naast het genereren van code kan generatieve AI ook helpen bij het maken van automatische documentatie. Doordat het AI-model de werking van de code begrijpt kan het een beschrijvingen genereren in een natuurlijke taal. Dit maakt de code niet alleen begrijpelijker, het maakt het ook mogelijk voor nieuwe programmeurs sneller de werking van bepaalde code te begrijpen.
- **User Interface (UI) ontwerp:** Generatieve AI kan worden ingezet om prototypes van UI-elementen te genereren op basis van ontwerpspecificaties. Dit versnelt het ontwerpproces en maakt het gemakkelijker om iteraties en aanpassingen aan te brengen.
- **Testautomatisering:** Generatieve AI kan testscripts genereren en uitvoeren, waardoor automatische testen effectiever worden. Dit verhoogt de testdekking en helpt bij het identificeren van potentiële problemen vroeg in het ontwikkelingsproces.

Deze toepassingen van generatieve AI hebben het potentieel om de efficiëntie, kwaliteit en snelheid van softwareontwikkeling drastisch te verbeteren. Het stelt ontwikkelaars in staat om zich te concentreren op creatieve en complexe taken, terwijl repetitieve en tijdrovende taken worden geautomatiseerd. Op deze manier wordt kostbare tijd van schaarse expertise bespaard en wellicht nog belangrijker, wordt het werk van een ontwikkelaar nog leuker.

De belofte en potentie is enorm, maar wat zijn de feitelijke resultaten? Op het moment van schrijven is er nog maar weinig meetbaar onderzoek gedaan naar de behaalde voordelen in projecten. De resultaten zijn lastig te valideren en moeilijk te vergelijken omdat de tools bijna dagelijks veranderen en het resultaat erg afhankelijk is van de context waarin het project is uitgevoerd. Ervaring met de tool, programmeertaal, ervaring van de ontwikkelaar, complexiteit van de opdracht, soort opdracht (frontend, backend of database), bestaande of nieuwe code. Het zijn factoren die allemaal impact hebben op het resultaat. Om de resultaten van verschillende projecten goed te kunnen vergelijken heeft Capgemini een 'measurement framework' opgezet zodat alle projecten die gebruik maken van Generatieve AI op dezelfde manier de resultaten meten. Deze centrale database kan in de toekomst gebruikt worden voor analyses en het maken van de business case.



DE WERKING VAN GENERATIEVE AI IN SOFTWAREONTWIKKELING

Generatieve AI-modellen maken gebruik van complexe neurale netwerken om grote hoeveelheden bestaande gegevens te analyseren. Hierdoor kunnen ze patronen en structuren in gegevens ontdekken die op basis daarvan nieuwe inhoud genereren. Naarmate het netwerk zelf groter en complexer wordt neemt ook de mogelijkheid om meer complexe opdrachten en output te genereren toe.

EEN KIJKJE ONDER DE MOTORKAP VAN GROTE TAALMODELLEN

In de trainingsfase wordt het model getraind op een enorme hoeveelheid tekstgegevens. Tijdens deze fase leert het model patronen en structuren in de tekst herkennen, zoals grammaticaregels, woordbetekenissen en contextuele relaties.

Grote taalmodellen zoals ChatGPT maken gebruik van een transformer-architectuur. Dit is een specifiek type neuraal netwerk dat uitstekend geschikt is voor het begrijpen van relaties tussen woorden en zinnen. Het bestaat uit meerdere lagen van neuronen die samenwerken om informatie te verwerken.

Tijdens de pre-training fase wordt het model “pre-trained” op de tekstgegevens, wat betekent dat het leert om automatisch zinnen aan te vullen (het zogenaamde “masked language modeling”) om contextuele representaties van woorden te creëren. Hierdoor begrijpt het model de relaties tussen woorden en zinnen.

Na de pre-training kan het model worden gefinetuned voor specifieke taken zodat het geoptimaliseerd kan worden voor bepaalde toepassing zoals het vertalen van tekst, samenvattingen maken, of code genereren.

Wanneer het model wordt gebruikt voor het genereren van tekst, zoals het beantwoorden van vragen, begint het met een initiële tekst of prompt. Het gebruikt de context van die tekst om vervolgens nieuwe tekst te genereren die past bij de gegeven input. Het doet dit door waarschijnlijkheidsberekeningen te maken op basis van wat het heeft geleerd tijdens de training.

Een voorbeeld:

We bouwen een interne applicatie voor Capgemini waarin we van consultants bijhouden welke expertise ze hebben, op welke opdracht en bij welke klant ze werken. Voor het bouwen van de database geven we ChatGPT de volgende opdracht (prompt).

You are a database engineer, can you write a SQL script that creates a database with the tables Consultants[name, role, city, rate], Expertise [name, level, certification], Clients[name, accountmanager], Assignment [startdate, enddate, project]. Consultants can have more than one expertise. Consultants can have one or more assignments. An assignment always has one client.

DH

ChatGPT is onder andere getraind om de taal Structured Query Language (SQL) te begrijpen. Het kan op basis van de prompt een script genereren die de juiste tabellen, kolommen en sleutels aanmaken. Op basis van de bovenstaande prompt heeft ChatGPT de volgende zaken begrepen:

- Voor de kolom ‘rate’ begrijpt het dat het een decimaal getal moet zijn en voor startdate en enddate begrijpt het dat het van het DATE type moet zijn. Voor de overige kolommen wordt er gebruik gemaakt van het type varchar. Daarnaast zijn alle tabellen voorzien van een sleutel (primary key).
- Consultants en Expertise is een zogenaamde veel-op-veel relatie (n:n) omdat één consultant meerdere expertises kan hebben, maar één expertise ook een relatie kan hebben met meerdere consultants. Een veel-op-veel relatie moet in een relationele database gemodelleerd worden middels een tussentabel. ChatGPT heeft dit op basis van de input begrepen en een tussentabel ConsultantsExpertise gemaakt. De sleutel (primary key) is een samengestelde sleutel bestaande uit de consultants_id en expertise_id. Dit is een best practice, maar merk op dat dit niet expliciet in de prompt is meegegeven.
- De Assignment tabel heeft een relatie met de tabel Client en Consultant. Hiervoor worden in het in het script de juiste Foreign Key relaties aangemaakt.

In de prompt is niet expliciet aangegeven voor welke database server het script geschikt moet zijn. Hieronder een deel van het SQL script dat ChatGPT heeft genereerd. Het ‘AUTO_INCREMENT’ keyword voor het sleutelveld onthult dat het script gegenereerd is voor MySQL databases.

```
..Create the Consultants table
CREATE TABLE Consultants (
  consultant_id INT PRIMARY KEY AUTO_INCREMENT,
  name VARCHAR(255) NOT NULL,
  role VARCHAR(255),
  city VARCHAR(255),
  rate DECIMAL(10, 2)
);
..Create the Expertise table
CREATE TABLE Expertise (
  expertise_id INT PRIMARY KEY AUTO_INCREMENT,
  name VARCHAR(255) NOT NULL,
  level VARCHAR(255),
  certification VARCHAR(255)
);
..Create a table to represent the
many-to-many relationship between Consult
CREATE TABLE ConsultantExpertise (
  consultant_id INT,
  expertise_id INT,
  PRIMARY KEY (consultant_id, expertise_id),
  FOREIGN KEY (consultant_id) REFERENCES Consultants(consultant_id),
  FOREIGN KEY (expertise_id) REFERENCES Expertise (expertise_id)
);
```

DE IMPACT VAN GENERATIEVE AI GAAT VERDER DAN CODING ASSISTANTS

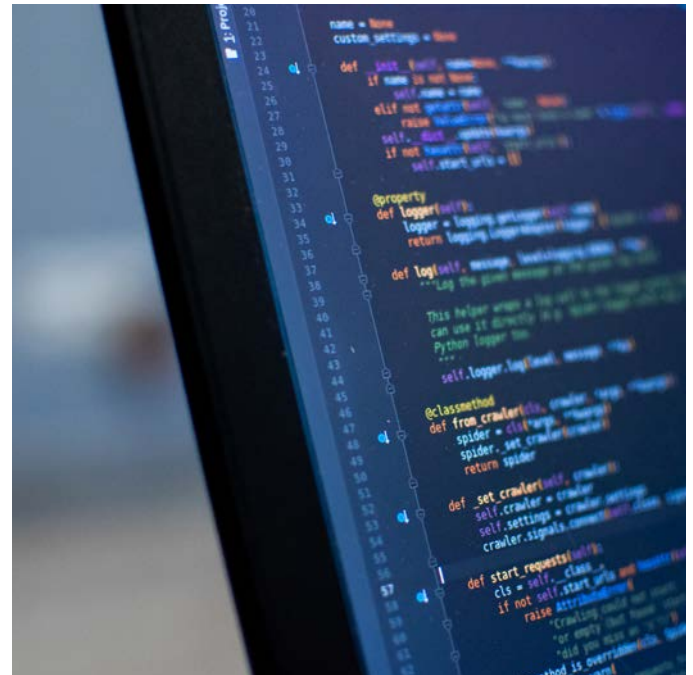
Het gebruik van generatieve AI overstijgt ruimschoots de rol van louter programmeerassistenten. Met gebruik van een natuurlijke taal (zoals Nederlands of Engels) kunnen ontwikkelaars de beoogde functionaliteit van nieuwe software beschrijven terwijl generatieve AI hun ideeën tot leven brengt. Andrej Karpathy, van OpenAI, trok recentelijk de aandacht door te zeggen: "De populairste nieuwe programmeertaal is Engels," waarbij hij het belang van deze doorbraak benadrukte. Generatieve AI kan worden geïntegreerd in de gehele levenscyclus van softwareontwikkeling (SDLC), van de analyse van zakelijke behoeften en het schrijven van user stories tot softwareontwerp, codering, documentatie, implementatie, testen en monitoring. Dit verrijkt de vele taken die softwareontwikkelaars uitvoeren.

Generatieve AI is één van de snelst groeiende technologieën die we ooit hebben gezien, en 96% van de CxO's zijn erop gebrand de mogelijkheden te verkennen en de voordelen te benutten² meer productiviteit, betere kwaliteit, versnelde time-to-value.

VERTROUWELIJKHEID EN INTELLECTUELE EIGENDOM (IP)

Het gebruik van generatieve AI kan leiden tot zorgen over vertrouwelijkheid en intellectueel eigendom. Invoer of opdrachten (zogenaamde prompts) die worden gegeven aan ChatGPT kunnen worden gebruikt om het model te trainen om zo optimalisaties aan te brengen in het model. Indien de opdrachten of invoer sensitive informatie bevat bestaat het risico dat deze informatie terecht komt in verbeterde versies van het model en zo onbedoeld in gegenereerde output van andere bedrijven of gebruikers terecht kan komen. Inmiddels hebben de meeste clouddiensten al een oplossing om taalmodellen zoals GPT-4 en PaLM 2 in een beveiligde omgeving aan te bieden. Hierdoor kunnen de data die gebruikt worden als input en training niet terecht komen in het publieke model en blijven de data binnen de eigen omgeving.

Een andere zorg die bedrijven hebben is het risico van inbreuk op intellectuele eigendomsrechten (IP) wanneer ze de resultaten gebruiken die worden gegenereerd door generatieve AI. Dit is begrijpelijk, gezien recente openbare onderzoeken door auteurs en kunstenaars naar hoe hun eigen werk wordt gebruikt in samenwerking met AI-modellen en -diensten. Deze bezorgdheid vormt mogelijk het voornaamste obstakel voor bedrijven die codegeneratie op basis van generatieve AI willen implementeren. Het positieve nieuws is dat cloudserviceproviders hier inmiddels oplossingen voor hebben aangekondigd of recentelijk hebben geïmplementeerd. Zo heeft Microsoft onlangs het Copilot Copyright Commitment geïntroduceerd. Hierbij neemt Microsoft de juridische risico's op zich in geval van mogelijke auteursrechtelijke geschillen. Ook Google onderkent het probleem en heeft maatregelen aangekondigd. Zo gaat Google Bard de bronvermelding toevoegen aan de gegenereerde code (output). Het doet dit op basis van de codepatronen en de bronnen die tijdens de training van het model zijn gebruikt.



Het gebruik van generatieve AI kan leiden tot zorgen over vertrouwelijkheid en intellectueel eigendom. Invoer of opdrachten (zogenaamde prompts) die worden gegeven aan ChatGPT kunnen worden gebruikt om het model te trainen om zo optimalisaties aan te brengen in het model.

Ontwikkelaars hebben nu de keuze om de gegenereerde code op te nemen in het project, of om het als open-sourceproject te implementeren met het voordeel dat dit project al onderhouden wordt door de community.

DE KOSTEN VAN GENERATIEVE AI

Hoewel generatieve AI de efficiëntie kan verbeteren, kunnen ongecontroleerde toepassingen leiden tot onverwachte kosten. In de toepassing van het genereren van programmeercode zullen ontwikkelaars voornamelijk gebruik maken van diensten zoals Github Copilot of Googles Bard. De modellen die deze diensten gebruiken zijn specifiek getraind met programmeercode. Voor deze diensten wordt veelal een vast maandelijks bedrag betaald.

Indien gebruik gemaakt wordt van één van de OpenAI modellen voor bijvoorbeeld het genereren van documentatie of user stories zullen de kosten bepaald worden op basis van het aantal 'tokens'. Als richtlijn kan je uitgaan van ongeveer 1000 tokens per

750 woorden. Als voorbeeld: voor de output die GPT-4 genereert kost het gebruik \$0.12 per 1000 tokens terwijl dit \$0.0016 is voor het eenvoudigere model babbage-002.

IMPACT OP SOFTWARE ENGINEERING

De adoptie van generatieve AI kan de structuur en werkwijze van software engineering teams veranderen. Ontwikkelaars kunnen zich meer richten op het formuleren van ideeën en het bepalen van de gewenste functionaliteit, terwijl AI verantwoordelijk is voor de feitelijke implementatie. Dit vereist mogelijk aanpassingen in teamvaardigheden en workflows.

Het begrijpen van deze uitdagingen en overwegingen is van cruciaal belang voor organisaties die generatieve AI willen integreren in hun softwareontwikkelingsproces. Het potentieel is enorm, maar het moet zorgvuldig worden beheerd en geïmplementeerd om de beloofde voordelen te realiseren en de mogelijke risico's te minimaliseren.

OVER DE AUTEUR



DONALD HESSING

Vice President | Head of Microsoft, CTO Cloud & Custom Applications

Donald geeft leiding aan het Microsoft cluster binnen de afdeling Cloud & Custom Applications. Daarnaast is Donald als CTO verantwoordelijk voor de visie en technologiestrategie van de afdeling. Hij is gespecialiseerd in software engineering, enterprise architecture en strategieontwikkeling van business tot technologie.

[in https://www.linkedin.com/in/donaldhessing](https://www.linkedin.com/in/donaldhessing)

[✉ donald.hessing@capgemini.com](mailto:donald.hessing@capgemini.com)

[✕ https://twitter.com/dhessing](https://twitter.com/dhessing)

09 SUSTAINABILITY BY DESIGN: EEN CRUCIALE NIEUWE NORM IN CLOUD-ARCHITECTUUR

KLIMAATVERANDERING? HET IS HOOG TIJD DAT WE SOFTWARE
GAAN BESCHOUWEN ALS DEEL VAN DE OPLOSSING – EN NIET
VAN HET PROBLEEM



HIGHLIGHTS

- De efficiëntie van hardware weet steeds minder gelijke tred te houden met de groeiende complexiteit van software en groeiende datavolumes. De emissies die het gevolg zijn van software nemen daardoor toe. Reductie van deze emissies wordt steeds belangrijker.
- Verduurzaming van de cloud is een gedeelde verantwoordelijkheid van de aanbieder van de cloud en de klant.
- Principes van groene software zijn gericht op reductie en niet op neutralisering. Het doel is om überhaupt geen koolstof uit te stoten.
- Door in te zetten op duurzaam ontwerp en duurzame ontwikkeling van de cloud, kunnen aanbieders blijven inzetten op effectieve dienstverlening en tegelijk de kritische toets doorstaan van overheden en maatschappij als het gaat om reductie van de impact op de omgeving.
- Organisaties kunnen profiteren van kostenoptimalisaties die hand in hand gaan met verduurzaming.

Volgens het World Economic Forum zitten we midden in de Vierde Industriële Revolutie (4IR). 4IR brengt allerlei belangrijke technologische ontwikkelingen met zich mee, bijvoorbeeld als het gaat om artificial intelligence, robotics, genomics, materialen en 3D-printen. Ondertussen verzamelen en analyseren bedrijven, overheden en publieke organisaties meer data dan ooit, en de schaal, snelheid en diepgang waarmee dat gebeurt neemt continu toe. Grote aanjager achter al deze ontwikkelingen en de adoptie ervan is de public cloud.

Er is nog een grote ontwikkeling: de steeds luider klinkende roep om verduurzaming, vanuit zowel maatschappij als overheid. Een roep die bijvoorbeeld tot uiting komt in de Europese Corporate Sustainability Reporting Directive (CSRD), die vereist dat grote bedrijven inzicht geven in hun impact op mens en planeet. De cloud kan ook hier van grote waarde zijn; niet alleen om emissies te monitoren, maar ook om koolstofemissies die het gevolg zijn van softwaregebruik en dataverwerking te meten en reduceren.

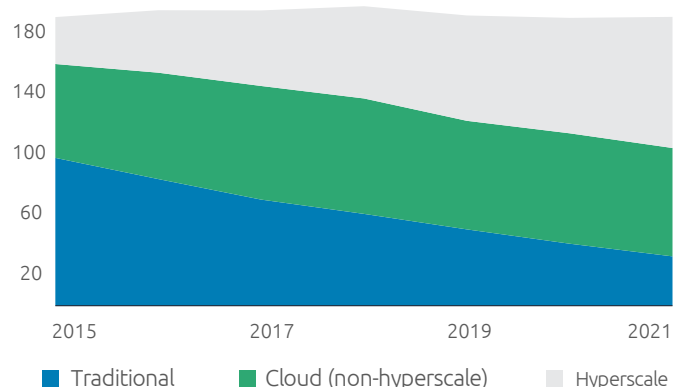
De Brundtland Commissie van de VN definieerde duurzaamheid in 1987 als volgt: 'meeting the needs of the present without compromising the ability of future generations to meet their own needs'.

Laten we beginnen met een nadere analyse van het woord sustainability, duurzaamheid. De Brundtland Commissie van de VN definieerde duurzaamheid in 1987 als volgt: 'meeting the needs of the present without compromising the ability of future generations to meet their own needs'.

WAT IS HET VERBAND TUSSEN SOFTWARE-ARCHITECTUUR EN DUURZAAMHEID?

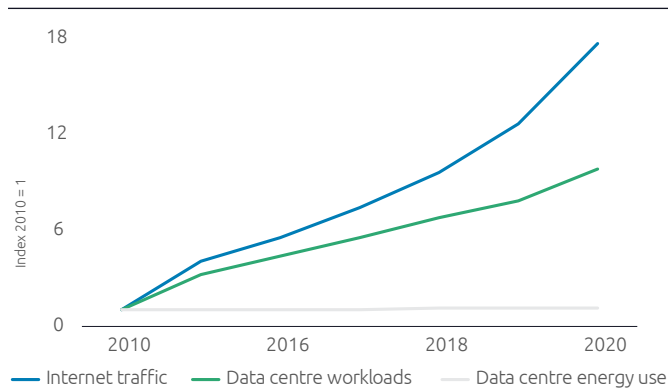
Software draaien betekent stroom verbruiken. Of het nu gaat om de infrastructuur waarop je software draait, de opslag van data of het versturen van data binnen het netwerk: het kost allemaal stroom. De energie die wordt verbruikt om hardware te produceren en te laten draaien is nog steeds voornamelijk afkomstig uit fossiele bronnen – met uitstoot van CO₂ als resultaat. Gelukkig worden servers, apparaten voor opslag en data center-infrastructuren steeds efficiënter. De kleine, inefficiënte datacentra van weleer worden steeds vaker vervangen door grotere, efficiëntere cloud-oplossingen, ondersteund door datacentra die al dan niet hyperscale zijn. Ontwikkelingen als deze hebben ervoor gezorgd dat de groei van de vraag naar elektriciteit het afgelopen decennium beperkt kon worden (zie Figuur 1 en Figuur 2), terwijl de workloads en het verkeer fors toenamen.

Figuur 1: Wereldwijde energievraag van datacentra, per type datacentrum



Source: IEA. Licence: CC BY 4.0. This data is subject to the IEA's terms and conditions: https://www.iea.org/t_c/termsandconditions/

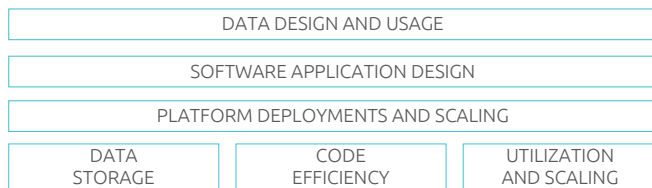
Figuur 2: Wereldwijde trends in internetverkeer, datacentrum-workloads en energievraag van datacentra



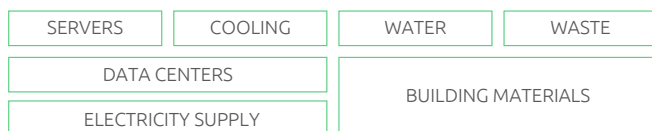
Source: IEA. Licence: CC BY 4.0

This data is subject to the IEA's terms and conditions:
https://www.iea.org/t_c/termsandconditions/

De consument is verantwoordelijk voor het verduurzamen van de workload binnen cloud



De hyperscaler is verantwoordelijk voor het verduurzamen van de infrastructuur binnen cloud



Het zal niet lang meer duren totdat verbeteringen aan hardware geen gelijke tred meer kunnen houden met toenemende complexiteit van data en toenemende datavolumes (Wirth's Law). De huidige staat van software en van data footprints is slechts een fractie van wat ons in het komende decennium staat te wachten. De energieconsumptie zal hierdoor toenemen, terwijl we maatregelen treffen om applicaties en databaseer efficiënter te maken.

'Duurzaamheid is een gedeelde verantwoordelijkheid.'

Hyperscalers als Microsoft Azure, AWS en Google Cloud hebben zichzelf doelen gesteld als het gaat om net-zero of zelfs negatieve emissies van koolstof. Microsoft wil in 2030 net-zero hebben bereikt, Amazon in 2040 en Google ook in 2030. Dat neemt niet weg dat klanten verantwoordelijk blijven voor de footprints in de cloud van hun applicaties en hun data. Het is dus ook aan hun om de workloads hiervan duurzamer te maken.

WAT IS EEN DUURZAME WORKLOAD?

"Een duurzame workload draait om oplossingen die zijn ontworpen voor maximale inzetbaarheid, terwijl ze zo weinig mogelijk verspilling opleveren. De footprint wordt hierdoor beperkt." - Microsoft

De impact en voordelen van applicaties die worden gehost in de cloud, werden altijd vastgesteld vanuit het perspectief van kosten en/of efficiency. Dat gold ook voor de Well Architected Frameworks (WAF) van hyperscalers als Azure en AWS. Volgens de laatste inzichten volstaat die nauwe blik niet meer. Een goed ontworpen cloudapplicatie moet ook duurzaam zijn. Azure¹ en AWS² hebben inmiddels hun WAF uitgebreid met dat duurzaamheids-perspectief. Hun frameworks bieden tegenwoordig richtlijnen voor het ontwerp van duurzame workloads. Daarbij volgen ze de Green Software Principles.

De Green Software-filosofie heeft als doel meer inzicht te creëren in wat er nodig is om oplossingen met een kleinere CO₂-footprint te ontwerpen en je CO₂-impact te meten, om uiteindelijk onnodig energieverbruik en onnodige emissies terug te dringen.

WAT IS GREEN SOFTWARE?

Green Software is software die minder broeikasgassen tot gevolg heeft. De door de Linux Foundation opgerichte Green Software Foundation³ is een ecosysteem van mensen, standaarden, tooling en best practices rond Green Software - gericht op echte reductie, en niet slechts op neutralisering.

Figuur 3: Shared responsibility model
 (Adapted from AWS Well Architected Framework)

GREEN SOFTWARE PRINCIPLES



Energy Efficiency

Use the least amount of energy possible



Hardware Efficiency

Use the least amount of embodied carbon⁴ possible



Carbon Awareness

Do more when the electricity is cleaner and do less when the electricity is dirtier

1. <https://learn.microsoft.com/en-us/azure/well-architected/sustainability/>
 2. <https://docs.aws.amazon.com/wellarchitected/latest/framework/sustainability.html>
 3. <https://greensoftware.foundation/>

4. Embodied Carbon: Embodied carbon (ook wel "embedded carbon" genoemd) is de hoeveelheid CO₂-verontreiniging die wordt uitgestoten tijdens de productie en de afvoer van hardware.

ASPECTEN VAN SOFTWARE-ONTWERP DIE VAN INVLOED ZIJN OP DUURZAAMHEID

1. Applicatie-ontwerp

Door container-architecturen, microservices of serverless concepten toe te passen, voorkom je dat beschikbare bronnen te zwaar worden belast, of juist niet volledig worden benut. Zo borg je dat applicaties precies de resources mogen gebruiken die ze nodig hebben. Auto scaling berekent de juiste requirements, waardoor de beschikbare resources optimaal worden benut en kosten worden bespaard. Door selectief om te gaan met de informatie die een applicatie moet genereren, of met de chattiness van een API, beperk je de work load. Taken die niet tijdgebonden zijn kunnen asynchroon worden uitgevoerd, om de belasting van de resources te balanceren. Door applicaties cloud native te ontwerpen, optimaliseer je de performance en beperk je kosten.

2. Applicatie platform

Cloud-aanbieders bieden de nieuwste, efficiency-verhogende functionaliteiten. Als het even kan moet je je applicatie- en data-opslag realiseren in datacentra dichtbij de klant. Zo beperk je netwerkverkeer en voorkom je dat slechts 1 instantie wordt geladen. Taken van grote omvang kun je agenderen op dagdelen met een lage CO2-intensiteit (demand shifting), of zelfs toewijzen aan regio's waar de CO2-intensiteit lager is (demand shaping). Om de efficiëntie te verhogen kun je managed services toepassen; spot instances helpen je de workloads te flexibiliseren. Ongebruikte resources moet je identificeren en verwijderen. En tot slot: right sizing van de workloads is cruciaal.

3. Opslag

Sla alleen op wat relevant is. Dat is hier de sleutel. Organisaties moeten investeren in heldere classificering van data en in helder beleid omtrent retentie. Ongebruikte of overtollige data moeten worden verwijderd – tenzij het behoud ervan vanwege compliance- of audit-redenen nodig is. Via datacompressie kun je opslag-requirements verlagen. Naast het volume is het ook van belang om een duidelijke hiërarchie te definiëren als het gaat om datatoegang, op basis van toegangspatronen. Data die het vaakst wordt benaderd breng je onder in hot storage. Dat maakt deze data makkelijk toegankelijk, zonder dat daar excessieve rekenkracht voor nodig is. Als je werkt met databases, kun je de performance van queries verbeteren door optimaal gebruik te maken van indexering en datapartities. Wordt een applicatie wereldwijd gebruikt? Zorg er dan voor dat je opslagdiensten gebruikt die ondersteuning bieden aan locatie-gevoelige data retrieval. Op die manier is data altijd dichtbij de locatie beschikbaar waar deze worden gebruikt. Zo beperk je de belasting van het netwerk.

4. Netwerken en connectiviteit

Voorkom onnodig verkeer op het netwerk. Maak gebruik van Content Delivery Networks (CDN) om veelgebruikte statische data dichtbij de gebruikers op te slaan. Pas goede caching-technieken toe (CDN of application layer) om netwerkverkeer

te beperken. Maak compressie van netwerkbestanden mogelijk. Bouw begrenzings in in het netwerk, zodat verplaatsingen van data zoveel mogelijk plaatsvinden binnen dezelfde cloud en, indien mogelijk, binnen dezelfde regio.

5. Operationele procedures

“Je kunt niet verbeteren wat je niet kunt meten.”

In het ultieme ontwerp voor een duurzame cloud computing workload bestaat volmaakte harmonie tussen geoptimaliseerde kosten, beperkte uitstoot van CO2-emissies en geoptimaliseerd energieverbruik. Vanaf het begin moeten organisaties specifieke doelen vaststellen, de metingen identificeren die de mate van succes moeten aangeven, en ervoor zorgen dat de gestelde doelen haalbaar zijn.

Beperking van kosten gaat vaak hand in hand met beperking van uitstoot. Met andere woorden: als je erin slaagt de kosten van de workload te optimaliseren, zet je een belangrijke stap richting een meer duurzame workload.

CLOUD EFFICIENCY= RESOURCE COST + CARBON EMISSIONS + ENERGY CONSUMED

RESOURCE COST	+	CARBON EMISSIONS	+	ENERGY CONSUMED
• Optimize resources • Architecture apps for low costs		• Optimize emissions • Green architectures		• Optimize energy used • Green architectures

Laten we eens kijken naar een – versimpeld – voorbeeld van een workload die draait op Microsoft Azure.

SCENARIO

Technische requirement: Draai elke dag een set van performance tests.

Business Requirement: Data moeten binnen de EU staan en de tests moeten elke nacht om 00:00 worden uitgevoerd.

Non-functionele requirements: de oplossing moet kosteneffectief en duurzaam zijn.

De architectuur dicteert dat deze oplossing een Virtual Machine nodig heeft om alle tools en tests uit te voeren. Metingen wijzen uit dat de oplossing 4 uur nodig heeft om de tests te voltooien.

KOSTENBEREKENING

Aanname: De Virtual Machine draait continu.

Resource	Kwantiteit	Gebruik	Kosten
Virtual Machine (2vCPUs; 8GB RAM)	1	730 uur	€147.23

Laten we nu een aantal maatregelen treffen om de oplossing qua kosten en duurzaamheid te optimaliseren.

Stap 1: Optimaliseer gebruik en beperk uitstoot

De tests draaien slechts vier uur per dag. Het is dus behoorlijk verspillend om de Virtual Machine de hele dag te laten draaien. De VM kan worden geprogrammeerd om op te starten, de tests te draaien en daarna te stoppen/zichzelf uit te schakelen. Aangenomen dat het gehele proces 5 uur duurt, kan het gebruik van elektriciteit flink beperkt worden. Dat resulteert in aanzienlijke kostenbesparingen en lagere CO2-emissies.

Resource	Kwantiteit	Gebruik	Kosten
Virtual Machine (2vCPUs; 8GB RAM)	1	50 uur	€30.25

Stap 2: Maak de oplossing duurzamer

Microsoft publiceert per datacentrum een factsheet over de duurzaamheid van de installatie. Deze informatie wijst in het geval van het Nederlandse datacentrum bijvoorbeeld uit dat 79% van de gebruikte stroom afkomstig is van hernieuwbare bronnen in de vorm van windmolens. Datacentra in Zweden worden volgens Microsoft volledig bediend door hernieuwbare bronnen; een belofte die elke dag weer wordt ingelost.

We kunnen er dus voor kiezen om de workload naar het Zweede datacentrum te verhuizen, om zo het gebruik van hernieuwbare bronnen te borgen. Op die manier bevinden de data zich bovendien nog steeds binnen de grenzen van de EU. De migratie heeft ook gevolgen voor de prijs.

Resource	Kwantiteit	Gebruik	Kosten
Virtual Machine (2vCPUs; 8GB RAM)	1	50 uur	€27.68

Het bovenstaande is een goed voorbeeld van hoe je door aan de 'kostenknop' te draaien, de oplossing kunt verduurzamen.

CONCLUSIE

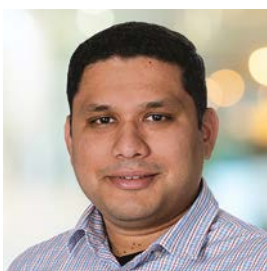
Door verspilling te beperken en je workload in de juiste regio in te richten, realiseer je een positieve impact op kosten en de CO2-footprint.

Het meten van CO2 kan uitdagend zijn. Zeker in de cloud. Organisaties maken vaak gebruik van het Greenhouse Gas (GHG)-protocol om hun totale CO2-emissies te meten. Binnen de cloud kan de scope van emissies die van toepassing is op een organisatie heel anders zijn dan de scope die van toepassing is op de cloud-hyperscaler. Er zijn tools beschikbaar (bijvoorbeeld Microsoft Sustainability Manager of Emissions Impact Dashboard, of de AWS Customer Carbon Footprint Tool) om gedurende langere tijd continu emissies te monitoren en veranderingen in emissies te analyseren. Op die manier kunnen organisaties betere beslissingen richting duurzaamheid nemen en voldoen aan publieke en externe rapportage-requirements.

Het is de komende jaren een voornaam doel voor CIO's en architecten: de integratie van duurzaamheid in de kern van software- en architectuur-ontwerp. 'Groene' software en groen beheer van data alleen zullen de klimaatverandering niet tot staan brengen, maar ze kunnen wel een belangrijke bijdrage leveren aan de wereldwijde reductie van CO2-emissies.

We raden je aan er nu mee te beginnen. Het zal nog een aantal jaren duren voordat duurzaamheid een basiselement zal vormen in ontwikkeling en in architectuur, maar tot dat moment telt elke besparing. Elke stap die we zetten richting de creatie van duurzame applicaties is een stap richting de invulling van wereldwijde klimaatdoelen. Laten we met elke actie die we ondernemen aan die doelen een bijdrage leveren. Laten we van duurzaamheid een requirement maken in onze ontwerpen, en niet slechts een bijzaak.

OVER DE AUTEUR



ARIN ROY

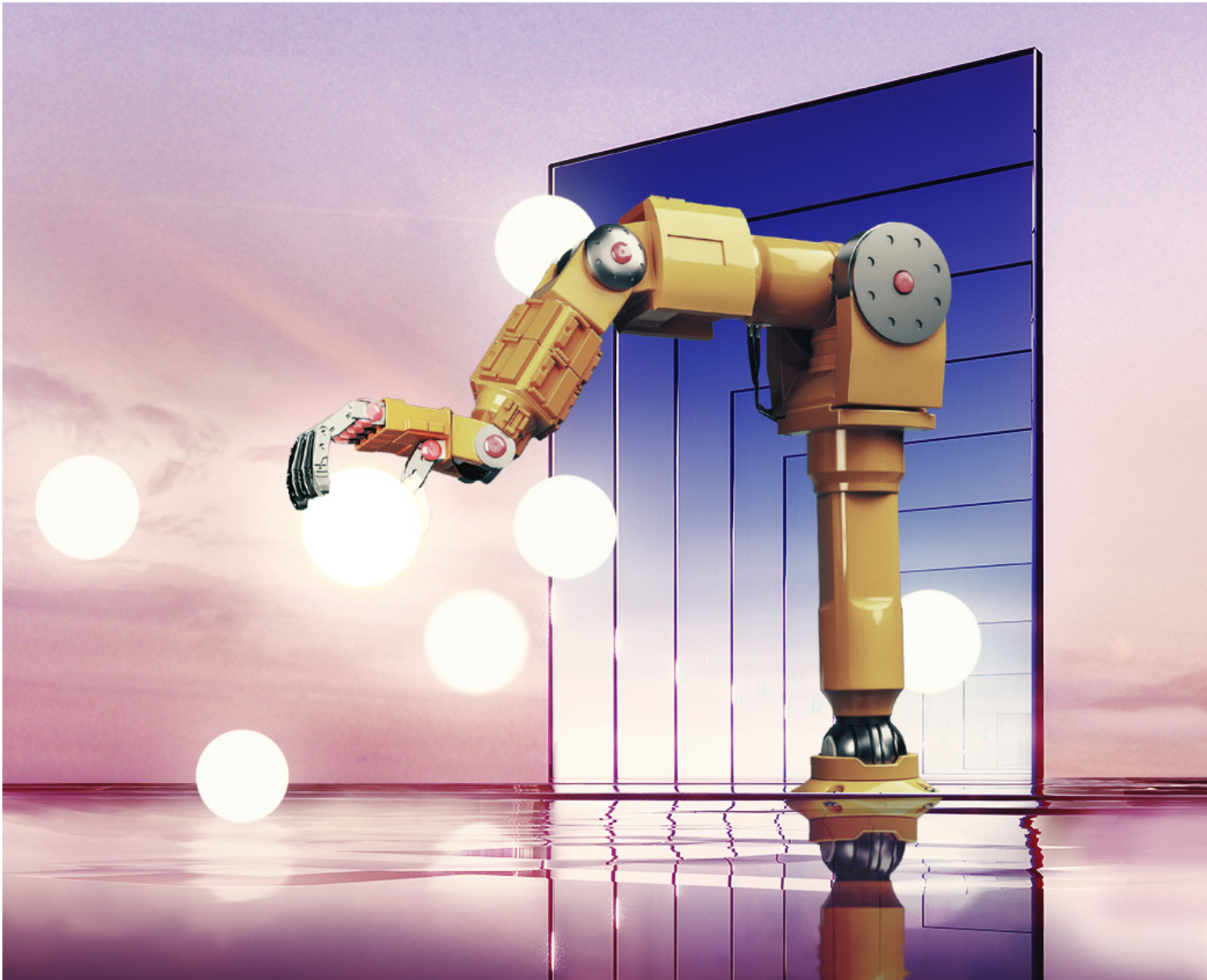
Tech Lead and Architect – Cloud Native

Als Tech Lead en Architect voor Cloud Native richt Arin zich op cloud native en identity solutions in Microsoft Azure. Arin heeft meer dan 15 jaar ervaring met de technologieën van Microsoft. Die ervaring helpt hem om enterprise-ready digitale cloud-oplossingen te ontwerpen en implementeren voor zijn klanten.

www.linkedin.com/in/arinroy

arin.roy@capgemini.com

<https://twitter.com/IamArinRoy>



10

MISSION IMPOSSIBLE: AUTOMATIC USABILITY TESTING

**DE LADYBUG: EEN TOOL VOOR SOFTWARETESTERS
OM SNELLER EN GESTRUCTUREERDER UX TESTEN
TE KUNNEN UITVOEREN**

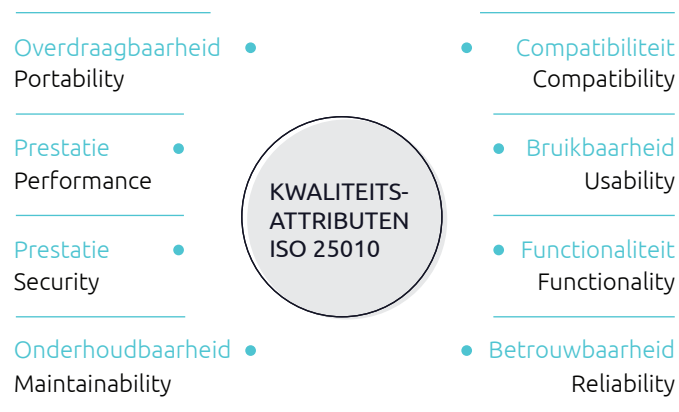
HIGHLIGHTS

- Uit een enquête ingevuld door meer dan honderd testers blijkt dat UX vaak niet gestructureerd wordt getest, maar door gebruik van intuïtie.
- De Ladybug tool zal op basis van de 10 usability heuristieken van Nielsen een deel van de UX automatisch kunnen testen.
- Aangezien UX gaat over de ervaring van de gebruiker blijft het cruciaal om eindgebruikers te betrekken bij de ontwikkeling en testen van een product. De Ladybug dient alleen ter ondersteuning van het testproces.
- Of heuristieken de beste manier zijn om UX te automatiseren valt nog over te betwisten, maar het is in ieder geval een start om testers naast het ondersteunen ook te informeren over het testen van UX.

Het voornaamste doel van softwaretesters in een ontwikkelteam is om te toetsen of de functionaliteit die gebouwd wordt voldoet aan de vooraf gestelde eisen. Daarbij wordt voornamelijk getest of het product 'doet wat het moet doen'.

Naast functionaliteit zijn er uiteraard ook andere 'kwaliteitsattributen' die getest kunnen worden, zoals beveiligbaarheid (security), prestatie (performance) en bruikbaarheid (usability) (Zie Figuur 1). Helaas raken deze non-functionele kwaliteitsattributen vaak ondergesneeuwd door andere taken die een tester moet doen, zoals testautomatisering. Vaak worden beveiligbaarheid en prestatie uitbesteed aan een gespecialiseerd team, maar bruikbaarheid lijkt vaak niet belangrijk genoeg gevonden om er serieus werk van te maken. Dat is jammer, want de mate van bruikbaarheid kan vaak hét verschil maken of een klant terugkeert op je website of het anderen aan zal bevelen.

Figuur 1: De 7 kwaliteitsattributen bij testen volgens ISO standaard 25010



HET PROBLEEM

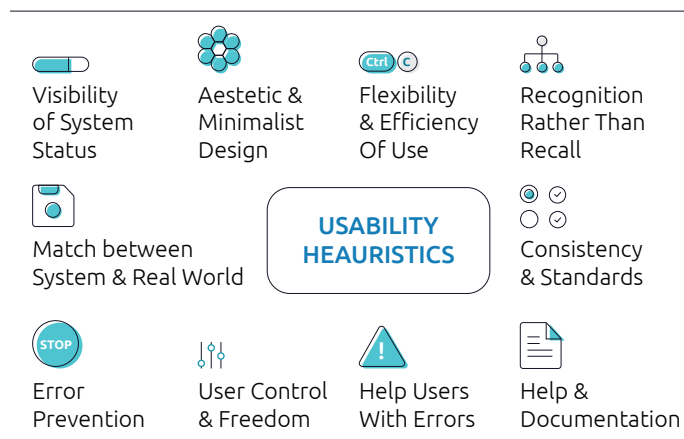
Waarom wordt de bruikbaarheid niet of nauwelijks getest? Uit gesprekken met collega's bleken de volgende zaken de belangrijkste bottlenecks:

1. Testers weten niet hoe ze UX **gestructureerd** moeten testen.
2. Agile testers ondervinden veel **tijdsdruk** en focussen daarom op de functionaliteit.
3. Er is een grote focus is op **testautomatisering**.
4. UX-bevindingen, waaronder bruikbaarheid, worden vaak niet hoog **geprioriteerd** door de business.

Met inachtneming van bovengenoemde bottlenecks zou het ideaal zijn om de bruikbaarheidstesten te automatiseren. Dat zou namelijk betekenen dat ze gestructureerd (bottleneck #1) en snel (bottleneck #2) kunnen plaatsvinden en uiteraard is bottleneck #3 hiermee ook gedekt. Echter, er wordt vaak

beweerd dat UX, waar bruikbaarheid een onderdeel van is, niet te automatiseren is omdat het subjectief is: het is de ervaring van de gebruiker, en deze verschilt dus ook per persoon. Er zijn daarentegen wel 10 Usability heuristieken van Nielsen (zie Figuur 2): vuistregels die al sinds 1990 bestaan en nog steeds gebruikt worden bij het ontwerpen van websites en applicaties.

Figuur 2: De 10 Usability Heuristieken van Nielsen¹



Er zijn inmiddels zelfs al een aantal tools beschikbaar die ontwerpers, ontwikkelaars en testers kunnen helpen om te testen aan de hand van deze 10 Usability Heuristieken, maar deze zijn nog niet geautomatiseerd². Deze vuistregels zijn uitgebreid beschreven op de website van de Nielsen and Norman Group (NNgroup) en met die informatie kan de tester aan de business uitleggen waarom een bruikbaarheidsbevinding zo belangrijk is, zodat deze beter geprioriteerd kan worden (bottleneck #4).

BEHOEFTE OM UX GESTRUCTUREERD TE TESTEN

Om te onderzoeken of bovenstaande bottlenecks correct zijn, maar ook of de testers gebruik zouden willen maken van een UX Automatiserings-tool is er een enquête opgezet. De enquête is ingevuld door 101 testers over de hele wereld. Dit aantal is te weinig om kwantitatief onderzoek te kunnen doen en statistisch gevalideerde conclusies te kunnen trekken, maar het geeft desalniettemin wel een goed beeld en bevestigt de benoemde bottlenecks³.

Uit de resultaten bleek dat ongeveer de helft van de testers de bruikbaarheid wel test en de andere helft niet of niet altijd. De testers gaven aan de bruikbaarheid niet (altijd) te testen omdat ze niet precies weten hoe ze dat moeten doen. Ook gaf 42% aan dat bruikbaarheid geen prioriteit is voor de business (werkende functionaliteit is voldoende).

Daarnaast bleek uit de enquête ook dat 97% misschien wel een tool zou willen gebruiken die hen kan ondersteunen bij het testen van UX testen.

1. <https://uxbooth.com/articles/newbie-heuristic-evaluation-mistakes-to-avoid/>

2. Chrome plugins Heurio en UXcheck

3. Lees het uitgebreide artikel over de enquête op: <https://huddle.eurostarsoftwaretesting.com/risingstar-update-is-there-a-need-for-a-ux-tester/> (Engels)

THE 'LADYBUG'

Het idee is dus om de bruikbaarheid van elke willekeurige website automatisch te laten testen door middel van de 10 usability heuristieken van Nielsen. Naast het automatiseren van een aantal tests zal de tool voornamelijk de tester informeren door middel van voorbeelden. De naam van deze automatische UX testing tool is: "Ladybug". Dit is de persona die jou aan de hand neemt bij het testen van de UX van de applicatie.

Een aantal voorbeelden:

AUTOMATISCHE TESTEN:

• Heuristiek #1: Zichtbaarheid van de status

De Ladybug tool zal de gehele website scannen op componenten. Als er helemaal geen spinner of loader op de website voorkomt zou je je kunnen afvragen of er echt nooit langer dan 1 seconde gewacht hoeft te worden op het laden van informatie. Als er namelijk geen indicatie is voor de gebruiker dat de applicatie bezig is, zou die zich af kunnen vragen of er wel iets gebeurt (en daardoor vaker op een knop drukken). In Figuur 3 zie je hoe de Ladybug response eruit zou kunnen zien.

Figuur 3: Ladybug response wanneer er geen spinner of loader gevonden is



Weet je zeker dat de gebruiker nooit langer dan 1 seconde hoeft te wachten op informatie de geladen wordt?

0.1 seconde is een directe reactie dus daar is er geen extra feedback voor nodig

1.0 seconde is acceptabel en er is geen feedback nodig zolang het onder 1 seconde blijft

10 seconden is te lang voor de gebruiker om gefocust te blijven op deze taak, dus is een indicatie van hoe lang het proces gaat duren belangrijk

Lees verder op <https://www.nngroup.com/articles/response-times-3-important-limits/>

• Heuristiek #9: Duidelijke foutmeldingen

Als de Ladybug op een 404 pagina geen link ziet in de content van de pagina, betekent het dat de gebruiker vast kan komen te zitten. Volgens Heuristiek #9 moet je niet alleen vertellen wanneer er iets mis is gegaan (404 error uitleggen), maar ook moet je de gebruiker helpen door een alternatief te bieden in de vorm van een link. In Figuur 4 is te zien hoe de Ladybug in dit geval zou kunnen reageren.

Figuur 4: Ladybug response op een 404 pagina zonder link



Ik heb geen linkje gevonden op de 404-pagina. Help gebruikers om uit deze foutpagina te komen.

Informeer gebruikers dat er iets mis ging

Vertel hen wat er mis ging

Reik hen een oplossing aan (bied een link aan)

Lees verder op <https://www.nngroup.com/articles/error-message-guidelines/>

• Heuristiek #10: Help en documentatie

Nog een check die gemakkelijk geautomatiseerd kan worden is de aanwezigheid van een FAQ-sectie op de website. Volgens Nielsen en Norman is het belangrijk voor verschillende soorten gebruikers om een plek te hebben waar ze naartoe kunnen om veel gestelde vragen te vinden. Dit geldt overigens niet voor alle gebruikers, maar voor degenen die dit prettig vinden is het fijn als deze sectie aanwezig is. In Figuur 5 kun je zien wat de Ladybug response eruit zou kunnen zien in het geval er geen FAQ sectie gevonden is.

Figuur 5: Ladybug response op een website zonder FAQ sectie



Ik heb geen FAQ gevonden in de footer van de pagina.

FREQUENTLY ASKED QUESTIONS

2 FREQUENTLY ASKED QUESTIONS

Goede FAQ pagina's zijn belangrijk voor gebruikers die extra informatie nodig hebben.

Lees verder op <https://www.nngroup.com/articles/error-message-guidelines/>

HANDMATIGE TESTEN

Er zullen altijd zaken blijven die niet automatisch kunnen worden getest. In die gevallen zal de Ladybug d.m.v. voorbeelden tips geven over hoe testers dit handmatig kunnen testen. In Figuur 6 zie je hoe de Ladybug advies kan geven over validatiemeldingen die niet automatisch getest kunnen worden.

Figuur 6: Het advies van Ladybug over validatiemeldingen.



Ik zie dat je keurig validatiemeldingen hebt bij de invoervelden. Zorg ervoor dat het duidelijk is wat de gebruiker moet doen om het probleem op te lossen. Dus niet "Dit veld is verplicht", maar "Het veld 'achternaam' mag niet leeg zijn".

ACHTERNAAM

Dit veld is verplicht

ACHTERNAAM

'Achternaam' mag niet leeg zijn



Lees verder op <https://wcag.nl/artikel/formulieren/>

RISINGSTAR2022, MAAR OOK KRITIEK

Het idee van het met een tool ondersteunen van UX testen heeft de RisingSTAR 2022 award gewonnen bij EuroSTAR: de grootste software testing conferentie in Europa. De prijs is gericht op het motiveren van jong talent, omdat er reeds veel erkenning is voor mensen die al lang in het vak zitten en veel hebben betekend voor de software testing wereld. Deze RisingSTAR prijs is bedoeld om net startende testers te motiveren.

Één van de punten van kritiek op deze tool was dat heuristieken eigenlijk niet bedoeld zijn om te automatiseren. De definitie van een heuristiek is: een vuistregel die wordt gebruikt om snel en met zo weinig mogelijk inspanning betekenis te geven aan een situatie. Dit is eigenlijk dus niet iets wat op wetenschap gebaseerd is, maar eerder door mensen zelf is bedacht op basis van eerdere ervaringen. Het is bij uitstek dus niet iets dat je zou willen automatiseren. Echter, het doel van de tool is: het ondersteunen van software testers en daarbij hen ook informeren over waar ze op moeten letten bij het testen van de bruikbaarheid. Voor dit doel zijn de heuristieken een mooi startpunt. Uiteindelijk zou de tool ook uitgebreid kunnen worden naar andere methodes om bruikbaarheid te kunnen testen. Er zijn namelijk ook andere heuristieken naast die van Nielsen. Het zou mooi zijn om hierin bijvoorbeeld ook toegankelijkheid (accessibility) mee te nemen. Toegankelijkheid en bruikbaarheid hebben namelijk veel raakvlakken er zijn nu al een aantal automatische tools om de toegankelijkheid te testen, zodat dit nu geen prioriteit hoeft te hebben. Uiteindelijk zou het echter goed zijn als er niet meerdere tools nodig zijn om de UX te testen.

Andere kritische vragen die naar aanleiding van de presentatie van het idee op EuroSTAR 2023 naar voren kwamen zijn:

1. Is het niet te laat als de tool werkt aan de hand van de URL van een website in productie? Moet het niet mogelijk zijn om tijdens het ontwikkelproces de UX te testen en niet alleen als de website al in productie is?
2. Hoe kunnen niet web-based applicaties getest worden op UX?

Dit zijn waardevolle punten waar goed over nagedacht moet worden. Eén van de mogelijkheden kan zijn om de tool ook lokaal te laten draaien.

VERVOLG

Inmiddels is ook de Europese Knowit Innovation Zone geïnteresseerd om bij te dragen aan de realisering van de Ladybug. Zij leveren gratis capaciteit voor innovatieve projecten. Het wordt een OpenSource project op GitHub, zodat iedereen die interesse heeft eraan kan bijdragen.

CONCLUSIE

De Ladybug kan zeker een toegevoegde waarde hebben om betere kwaliteit te kunnen leveren. Kwaliteit gaat namelijk verder dan alleen werkende functionaliteit, en bruikbaarheid is juist dát stapje extra waardoor gebruikers jouw product kiezen boven een ander.

Verder moet er nog onderzoek gedaan worden naar andere manieren, naast de heuristieken, om de bruikbaarheid en andere onderdelen van UX automatisch te kunnen testen. De Heuristieken dekken alleen de bruikbaarheid af, maar nog geen andere elementen van UX, zoals toegankelijkheid.

OVER DE AUTEUR



KIMBERLY SNOYL

UX Designer & Sr. Software Tester

Kimberly beschouwt het als haar persoonlijke missie om het belang van UX Testing kenbaar te maken onder software testers. Ze geeft sinds 2020 presentaties over dit onderwerp op verschillende (internationale) conferenties. In haar huidige rol zorgt zij ervoor dat de te ontwikkelen IT-oplossing voortdurend wordt getoetst met de eindgebruiker om ervoor te zorgen dat de applicatie aansluit op hun werkwijze en behoeften.

[in https://www.linkedin.com/in/kimberlysnoyl/](https://www.linkedin.com/in/kimberlysnoyl/)

[✉ kimberly.snoyl@capgemini.com](mailto:kimberly.snoyl@capgemini.com)

Previous blogs on the subject:

- <https://huddle.eurostarsoftwaretesting.com/risingstar-2022-update-mission-impossible-automatic-usability-testing/>
- <https://huddle.eurostarsoftwaretesting.com/risingstar-update-is-there-a-need-for-a-ux-tester/>

11

PRODUCT DISCOVERY: ECHTE WAARDE CREËREN IN PLAATS VAN ONGEBRUIKTE FEATURES BOUWEN

IN DELIVERY ONTWIKKELEN WE HET PRODUCT OP DE JUISTE MANIER, IN
DISCOVERY ZORGEN WE DAT WE HET JUISTE PRODUCT ONTWIKKELEN



HIGHLIGHTS

- Start met het delen van problemen, in plaats van oplossingen.
- Ontdekken of een nieuwe feature goed aansluit bij de behoeften van gebruikers is tijdens Product Discovery veel goedkoper dan na Product Delivery.
- In de toekomst vragen we aan teams vaker om problemen op te lossen, in plaats van features te bouwen.
- Stop met het meten van succes aan de hand van output, start met het meten van outcomes.
- Dual-Track agile combineert Product Discovery en Product Delivery voor een continue feedback loop.

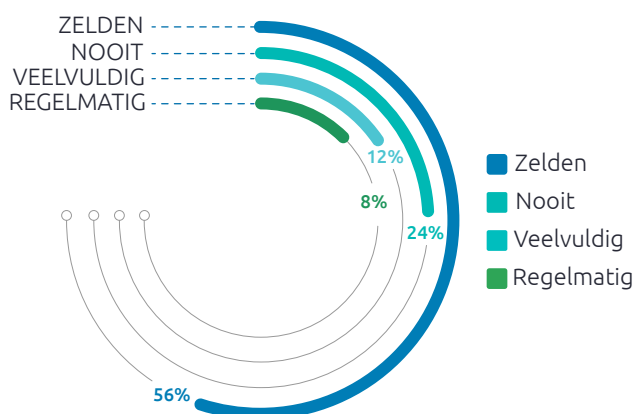
Hoe zeker zijn we dat de feature die we willen bouwen ook daadwerkelijk gebruikt zal worden? Kunnen gebruikers er makkelijk mee overweg? Sluit een nieuw idee wel echt aan bij de behoeften van gebruikers? Dit zijn vragen die teams met Product Discovery proberen te beantwoorden.

Met Product Discovery streven teams ernaar gebruikers beter te begrijpen en ideeën vooraf te valideren. Dit is essentieel, want 80% van de features wordt zelden tot nooit gebruikt.

Uit onderzoek blijkt dat 56% van alle features zelden gebruikt wordt en 24% zelfs nooit.

Hoe zeker zijn we dat de feature die we willen bouwen ook daadwerkelijk gebruikt zal worden? Kunnen gebruikers er makkelijk mee overweg? Sluit een nieuw idee wel echt aan bij de behoeften van gebruikers? Dit zijn vragen die teams met Product Discovery proberen te beantwoorden.

Figuur 1: Feature adoption report



Binnen veel organisaties bepaalt de Product Owner samen met gebruikers en stakeholders wát er gebouwd wordt, terwijl het development team bepaalt hóe het gebouwd wordt. Zodra een nieuwe feature gereed is en opgeleverd, ontvangen we feedback van gebruikers en stakeholders. Deze werkwijze kent echter twee belangrijke nadelen.

Ten eerste beschrijft de Product Owner, in samenspraak met gebruikers en stakeholders, niet alleen een kans, behoefte of probleem, maar bepalen zij vaak ook voor een groot deel ook al de oplossing. En dat is helaas niet altijd de beste oplossing. Gebruikers en stakeholders weten vaak pas echt wat ze willen nadat ze een oplossing hebben gezien. Bij Product Discovery is het hele team betrokken bij het vinden van de beste oplossing.

Het tweede nadeel van deze werkwijze is dat we pas ontdekken of de oplossing effectief is nadat we deze in productie hebben opgeleverd aan gebruikers en stakeholders. Het is verspilling van tijd en geld als de feature daarna niet gebruikt wordt of als er nog iteraties nodig zijn om het te verbeteren. Met Product Discovery valideren we de oplossing eerst, alvorens deze te realiseren.

Figuur 2: Product discovery en delivery



EERST HET PROBLEEM, DAN DE OPLOSSING

Product Discovery helpt teams bij het verkennen en valideren van ideeën om tot innovatieve oplossingen te komen die goed aansluiten bij de behoeften van gebruikers. Hierbij maken we onderscheid tussen het probleemdomein en het oplossingsdomein. Deze scheidslijn helpt ons om eerst goed de behoeften van gebruikers te begrijpen voordat we over gaan tot bedenken van oplossingen.

In het probleemdomein richten we ons op het identificeren en begrijpen van gebruikersproblemen, behoeften en kansen. Het doel hierbij is om diepgaand inzicht te krijgen in wat de gebruiker nodig heeft of wat ontbreekt in de huidige oplossing. Eén van de belangrijkste voordelen van het focussen op het probleemdomein is er zeker van te zijn dat er daadwerkelijk een probleem of behoefte bestaat voordat we over oplossingen gaan nadenken.

CASE STUDY: HET VERKENNEN VAN HET PROBLEEM- EN OPLOSSINGSDOMEIN BIJ EEN E-COMMERCEPLATFORM

Bij een e-commerceplatform kwam een stakeholder met het idee om een sectie 'recent bekeken producten' toe te voegen aan de site, in de hoop de conversieratio te verhogen. Naar aanleiding van dit voorstel besloot het team eerst onderzoek te verrichten. Ze analyseerden het gedrag van bezoekers die wel producten bekeken, maar geen aankoop deden. Daarnaast voerden ze interviews met personen uit deze groep. Uit het onderzoek bleek dat deze bezoekers voornamelijk moeite hadden met het vinden van het gewenste product en daarom meerdere producten bekeken. Dit waardevolle inzicht hielp het team om een daadwerkelijk probleem te identificeren dat het waard was om op te lossen.

Na het identificeren van het probleem zocht het team contact met een ander team binnen de organisatie dat verantwoordelijk was voor het zoekalgoritme. Het idee was om gezamenlijk te werken aan het verbeteren van de zoekfunctionaliteit. Dit omvatte zowel het verfijnen van het algoritme als het optimaliseren van de weergave van zoekresultaten.

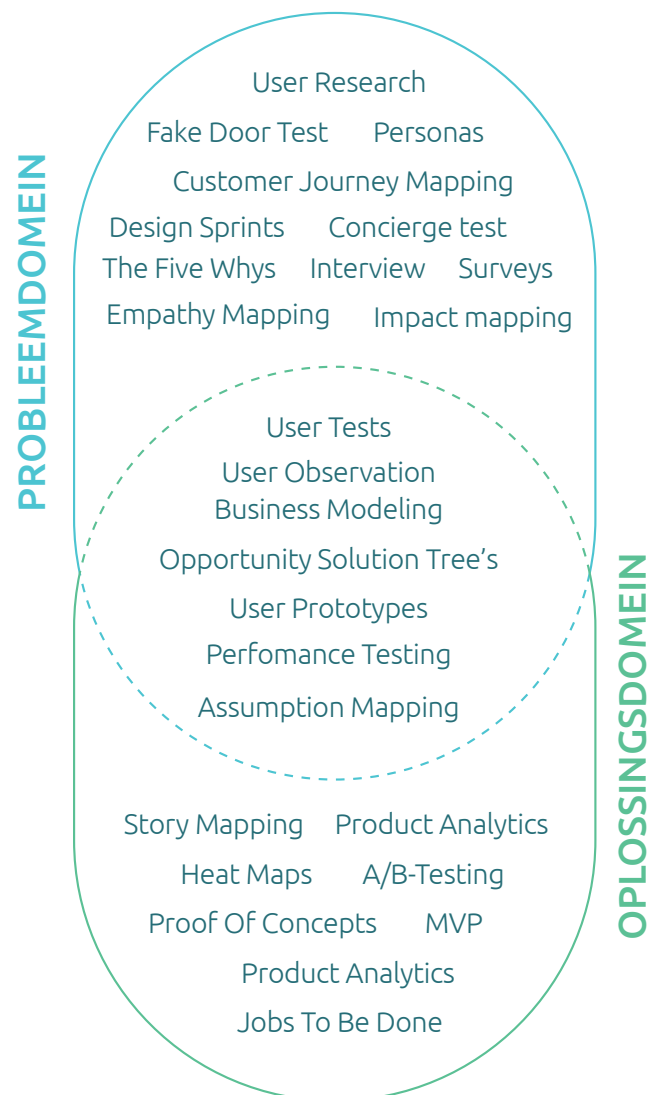
Om de effectiviteit van de wijzigingen te toetsen zonder al te grote risico's, besloten de teams om hun aanpassingen eerst uit te rollen naar een beperkte groep bezoekers. Dit stelde hen in staat om waardevolle feedback te verzamelen en de aanpassingen bij te schaven voordat ze grootschalig werden geïmplementeerd.

Tijdens dit proces kwam één van de developers met een innovatief idee: wat als ze, naast de directe zoekresultaten, ook gerelateerde producten zouden tonen? Deze gerelateerde producten zouden gebaseerd zijn op items die vaak samen werden gekocht. Dit idee leek kansrijk en werd opgenomen in de testfase.

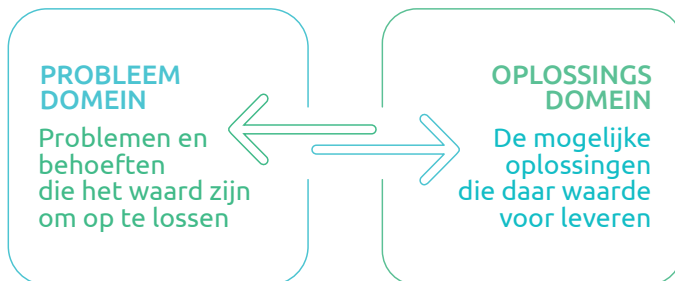
De resultaten waren veelbelovend. Het vernieuwde zoekalgoritme, gecombineerd met de weergave van gerelateerde producten, leidde tot een hogere conversieratio. Bezoekers kochten niet alleen de producten die ze oorspronkelijk zochten, maar in sommige gevallen ook de aanvullende, gerelateerde producten die nu in de zoekresultaten werden getoond.

In het oplossingsdomein richten teams zich op het bedenken, vergelijken en valideren van oplossingen. Het doel hierbij is om op een snelle en kostenefficiënte manier te toetsen of een voorgestelde oplossing het probleem daadwerkelijk oplost. Op basis van feedback itereren en verfijnen teams de oplossingen verder.

Figuur 3: Product Discovery technieken in het probleem- en oplossingsdomein



Door het probleem- en oplossingsdomein van elkaar te scheiden vermijden teams de valkuil van het springen naar oplossingen zonder eerst goed te begrijpen wat het werkelijke probleem of de echte kans is.



EERST INVESTEREN IN WAT WE MAKEN, DAARNA IN HOE WE HET MAKEN

Met het toepassen van Product Discovery leren teams veel eerder en sneller of een oplossing de gewenste uitkomst heeft. Als teams dit pas leren nadat een nieuwe feature is opgeleverd aan gebruikers, bijvoorbeeld na één of meerdere sprints, zijn de kosten om deze aan te passen vaak hoger, omdat de investering dan al is gedaan om de oplossing of losse feature met productiekwaliteit op te leveren.

Eén van de thought-leaders binnen Product Discovery, Marty Cagan, beschrijft dat we met Product Discovery op vier grote risico's focussen.

De vier risico's:

- 1. Waarde:** Dit risico heeft te maken met de vraag of de oplossing daadwerkelijk waarde zal toevoegen voor gebruikers en de organisatie. Werken we aan een kans of probleem waar onze gebruikers om geven?
- 2. Bruikbaarheid:** Zijn gebruikers in staat om het product te gebruiken? Is het gebruikersvriendelijk en gaat het gebruikt worden zoals wij verwachten?
- 3. Haalbaarheid:** Is de oplossing technisch realiseerbaar met de beschikbare technologieën en binnen de gegeven tijd en budget?
- 4. Levensvatbaar:** Vanuit een zakelijk perspectief bezien of het ook economisch zinvol is om de oplossing te realiseren en in de toekomst te blijven onderhouden.

Er wordt soms ook gesproken over een vijfde risico, namelijk of het ethisch verantwoord is. Zeker met de opkomst van machine learning en big data waarbij grote hoeveelheden persoonsgegevens verwerkt kunnen worden.

VAN OUTPUTS NAAR OUTCOMES

In Product Discovery draait het om het oplossen van problemen of het benutten van kansen in plaats van simpelweg features te bouwen. We spreken in Product Discovery daarom ook van een verschuiving van het werken aan outputs naar het behalen

outcomes. Output refereert naar de dingen die gemaakt worden, zoals een nieuwe feature in een applicatie of het ontwerp van een microservice. Outcomes zijn de veranderingen die outputs wel of niet teweegbrengen, zoals het verhogen van de conversieratio met 3%. Teams meten dan ook hun succes binnen Product Discovery aan het behalen van outcomes in plaats van het opleveren van outputs. Ze zijn pas succesvol als ze het probleem hebben opgelost.

Door met meetbare doelen te werken, kan een team duidelijk bepalen wanneer een oplossing succesvol is. Een manier om dit effectief te doen is door het gebruik van OKR's (Objectives and Key Results). Dit betekent dat de outcomes, zoals het verkorten van doorlooptijden of het verhogen van de gebruikerstevredenheid, gekwantificeerd en getrackt kunnen worden met behulp van deze methodiek. Meetbare doelen bieden dus duidelijk criteria voor succes en stellen teams in staat om oplossingen continu te evalueren en te verbeteren.

De manier van werken binnen teams is geëvolueerd

Binnen Product Discovery is de manier van werken binnen teams geëvolueerd. In plaats van in fases te werken en taken aan elkaar over te dragen, werkt het hele team samen in zowel



Door met meetbare doelen te werken, kan een team duidelijk bepalen wanneer een oplossing succesvol is. Een manier om dit effectief te doen is door het gebruik van OKR's (Objectives and Key Results).

Discovery als Delivery. Teresa Torres, een Product Discovery coach, introduceerde het 'Product Trio' concept, een nauwe samenwerking tussen Product Management, Design en Engineering. Deze drie rollen samen zijn verantwoordelijk voor het behalen van outcomes. Dit doen ze door van begin tot eind samen te werken. Van het spreken met gebruikers en stakeholders, tot het valideren van prototypes en het realiseren van de oplossing.

COMPETENTIES EN VAARDIGHEDEN IN PRODUCT DISCOVERY

Om effectief aan Product Discovery te doen, moeten teams direct contact onderhouden met gebruikers en stakeholders, bekwaam zijn in interviewtechnieken, en kunnen werken met designsoftware. Ze moeten hypothesen kunnen testen, risico's herkennen, technologische beperkingen begrijpen, en een sterke data-analyse vaardigheid bezitten voor datagedreven besluitvorming.

Het combineren van Product Discovery & Delivery

Het combineren van Product Discovery en Product Delivery door teams wordt ook wel Dual-Track agile genoemd. In deze methode lopen de Discovery track en Delivery track parallel aan elkaar. Dit betekent dat teams hierin continu bezig zijn met

het ontdekken, definiëren, en valideren van ideeën, terwijl ze tegelijkertijd werken aan het opleveren van nieuwe oplossingen. De inzichten die worden opgedaan nadat een oplossing is opgeleverd aan gebruikers kunnen weer terugvloeien naar de Discovery track als bron voor nieuwe ideeën of het verder verfijnen van een oplossing.

CONCLUSIE

Product Discovery is onmisbaar om op efficiënte wijze waardevolle en gebruikersvriendelijke oplossingen te ontwikkelen die goed aansluiten bij de behoeften van gebruikers. Het biedt een gefundeerde basis om te leren en te valideren waardoor de kosten voor het aanpassen van oplossingen drastisch worden verminderd. De methodologie benadrukt het belang van samenwerken door Product Management, Design en Engineering. Teams moeten zich richten op het oplossen van problemen en het behalen van meetbare doelen en niet langer enkel op het opleveren van features. Datagedreven besluitvorming is cruciaal; het omzetten van data in inzichten en acties. Om de vruchten te plukken van Product Discovery, dienen organisaties een cultuur van continue leren, validatie en samenwerking te omarmen en te investeren in vaardigheden zoals gebruikersinterviews, prototype-ontwerp en data-analyse. Vergeet niet, het gaat niet alleen om hoe we het bouwen, maar vooral om wát we gaan bouwen.

OVER DE AUTEUR



PHILIP WAGENAAR

Product Manager & Business Analyst

Philip heeft meer dan 20 jaar ervaring in product discovery en product delivery. In zijn huidige rol helpt hij organisaties in het domein van openbare orde en veiligheid in de ontwikkeling van technologie gedreven producten.

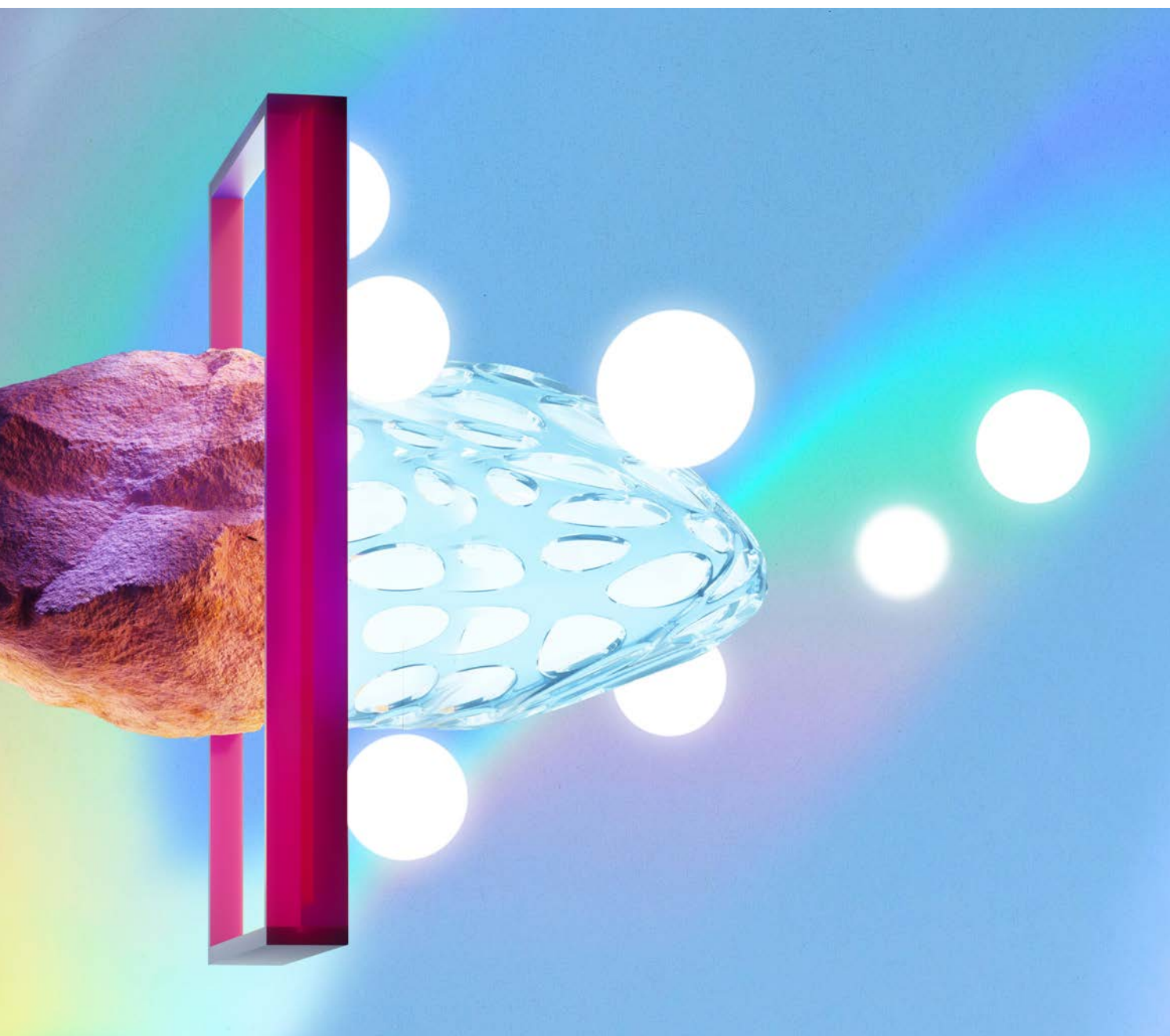
[in https://www.linkedin.com/in/philipwagenaar/](https://www.linkedin.com/in/philipwagenaar/)

[✉ philip.wagenaar@capgemini.com](mailto:philip.wagenaar@capgemini.com)

12

PLANT SIMULATION UNLEASHED: DE SLEUTEL TOT MANUFACTURING EXCELLENCE EN IT-SUPPORT

OPTIMALISATIE VAN OPERATIONS EN VERBETERING VAN BESLUITVORMING
IN DE PRODUCTIE DOOR MIDDEL VAN SIMULATIE TOOLS



HIGHLIGHTS

- Plant simulation biedt een dynamische oplossing om door de complexiteit van productieomgevingen te navigeren.
- Simulatiemodellering richt zich op de fabrieks lay-out, output planning, voorraadbeheer en productieplanning.
- Voorbeelden uit de praktijk tonen aan hoe simulatie tools efficiëntie, output en strategische besluitvorming bevorderen.
- Plant simulation evalueert de integratie van IoT-apparaten en sensoren in IT-systemen voor productie. Een gestructureerd 5-stappenkader helpt bedrijven om productieprocessen, ontwerpkeuzes en strategische beslissingen te voorspellen, optimaliseren en beoordelen.
- Simulatiemethoden zoals Agent-Based Modelling, Discrete-Event Modelling en Hybrid Modelling zijn geschikt voor uiteenlopende productiescenario's.
- Het legt de basis voor het creëren van digital twins, wat real-time monitoring en optimalisatie via IT-systemen mogelijk maakt.

In de huidige, turbulente productiewereld met scherpe klanteneisen, onzekerheid van aanvoer en complexe planningstrategieën, biedt plant simulation een effectieve aanpak om deze uitdagingen te lijf te gaan. In dit artikel bespreken we onze eigen ervaringen en verkennen we real-world toepassingen.

PRAKTISCHE TOEPASSINGEN VAN SIMULATIE SOFTWARE

Plant simulation gebruiken we met een eenvoudig doel: fabrikanten helpen om hun activiteiten efficiënter te laten verlopen. Wij werkten samen met een autofabrikant die zichzelf tot doel had gesteld de productie van auto's op te schroeven om aan de groeiende vraag te voldoen. De uitdaging daarbij was om de productie te verhogen zonder verstoringen te veroorzaken. In dit project werd AnyLogic als simulatiesoftware gebruikt. Eerst werd een basismodel met vaste parameters opgezet. Daarna werden verschillende configuraties gemaakt om te simuleren hoe de KPI's (Leverbetrouwbaarheid, utalisatie, vertraging en output) werden beïnvloed. Ten slotte werden de verschillende parameters gewijzigd om de geselecteerde KPI's te optimaliseren. De resultaten van deze simulaties gaven de fabriek een blauwdruk om hun output te verhogen.

Een ander voorbeeld van een simulatieproject was een wereldwijd chemisch bedrijf in Nederland. Het doel van het project was om de prestaties van alternatieve toeleveringsketenconfiguraties te evalueren op basis van de vastgestelde KPI's ("on time-in full", reactor utalisatie en voorraadniveau). "On time-in full" (OTIF) vertegenwoordigt de klantorder die volledig is ingevuld met de vooraf overeengekomen hoeveelheid. De uitdaging voor de fabriek was om een hogere OTIF te behalen en tegelijkertijd het voorraadniveau op een gewenst niveau te houden. In het

basismodel werd de huidige toeleveringsketenconfiguratie gemodelleerd in de software Tecnomatix Plant Simulation en dit model werd gebruikt voor validatie. Vervolgens werden alternatieve configuraties gemodelleerd in de simulatiesoftware en werden de prestaties van deze scenario's vergeleken met het basismodel. Dit hielp de fabriek om de knelpunten in de huidige situatie te analyseren en de best presterende (in relatie tot de KPI's) alternatieve configuratie werd gekozen voor implementatie.

REËLE TOEPASSINGEN

Met deze real-world voorbeelden in gedachten, leggen we uit hoe simulaties gebruikt kunnen worden voor verschillende toepassingen. De simulatieomgeving geeft ons ten minste twee use cases waar simulatie nuttig is:

- a) Het stelt ons in staat om de huidige, al bestaande processen te optimaliseren.
- b) Het stelt ons in staat om de ontwerpkeuzes voor een conceptuele fabriek of processen te evalueren.

Simulatie van de productie kan helpen om de volgende aspecten te evalueren, onder andere:

- Indeling van de fabriek en toewijzing van hulpbronnen
- Doorvoer- en capaciteitsplanning
- Beheer van voorraadniveaus en aanvuelsnelheden
- Productieplanning en de planning en balanceren van productielijnen
- Productielogistiek en materiaalstroom
- Procesverbetering, bottleneck analyse
- Predictive maintenance

Naast het optimaliseren van productieprocessen kan plant simulation ook IT-gerelateerde verbeteringen bieden die de algemene productieactiviteiten ondersteunen en verbeteren.

- Het optimaliseren van de IT-infrastructuur binnen de productiefaciliteit, om connectiviteit en datastromen tussen machines en systemen te garanderen.
- Het integreren van IoT-apparaten en -sensoren in de productieomgeving.
- Het creëren van een digital twin, wat real-time monitoring en optimalisatie via IT-systemen mogelijk maakt.

EEN FRAMEWORK VOOR SIMULATIE

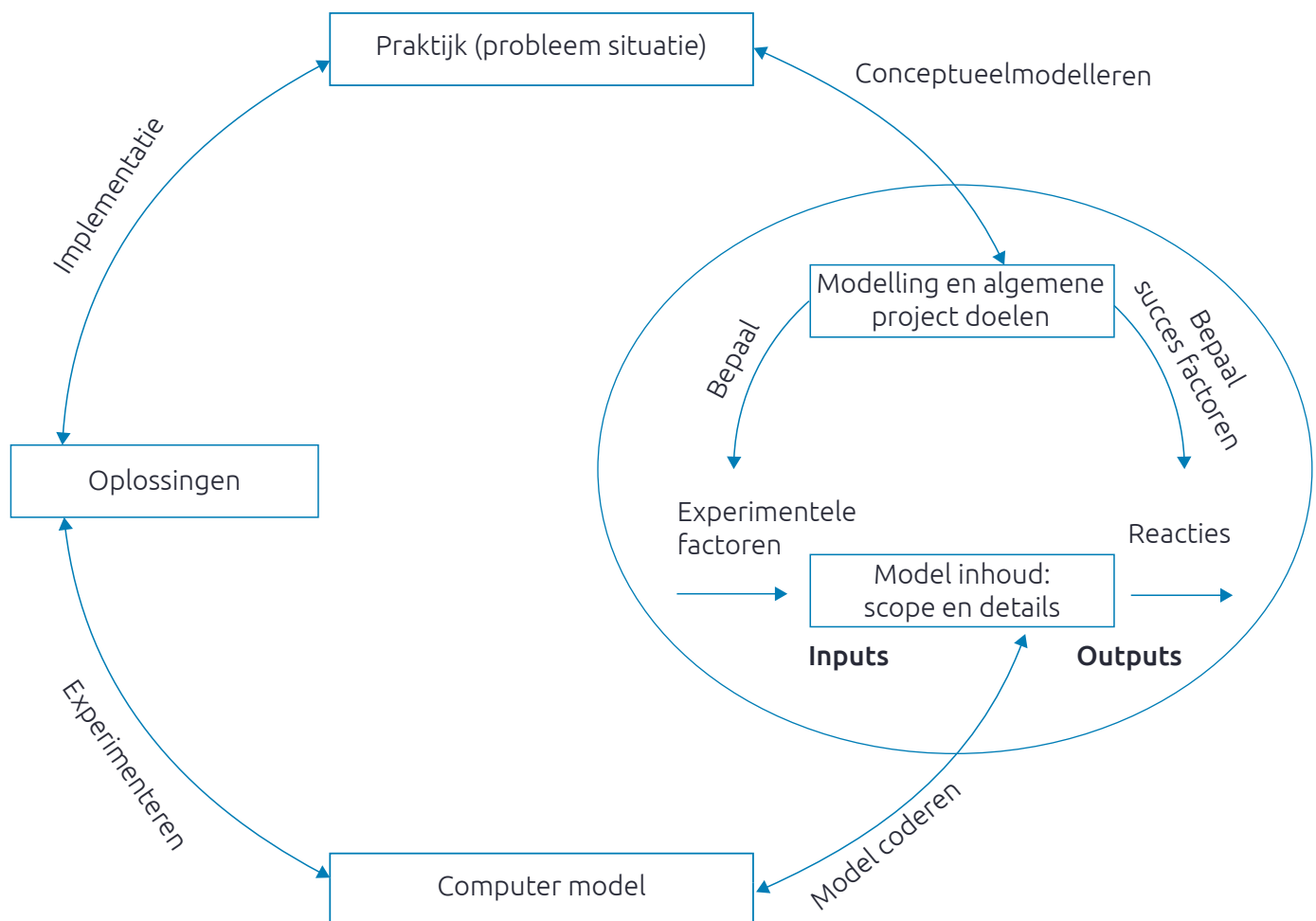
Succesvolle simulatieprojecten volgen een goed gedefinieerde aanpak. Door een 5-stappen benadering te gebruiken voor het voorbereiden, opzetten en gebruiken van simulatie, kan men snel een nieuwe of een bestaande productieomgeving evalueren en voorgestelde wijzigingen en alternatieve ontwerpen vergelijken.

In Figuur 1 geeft Robinson (2008)¹ een duidelijk kader van hoe een simulatie wordt gebouwd. In dit kader zijn er 5 stappen om een simulatie voor te bereiden, op te zetten en te gebruiken:

1. Probleemanalyse: Begin met een grondige analyse van het echte probleem dat u wilt oplossen.
2. Modelleringsdoelstellingen: Definieer duidelijk de doelstellingen van uw simulatie door middel van conceptuele modellering.

3. Omvang en doelen: Bepaal de omvang, het detailniveau en de doelen van de simulatie.
4. Modelontwikkeling: Ontwikkel een computermodel dat het systeem nauwkeurig weergeeft.
5. Implementatie van oplossingen: Haal oplossingen uit het computermodel en implementeer ze in de echte wereld.

Figuur 1 - Conceptual modelling for simulation part i: Definition and requirements (S.Robinson, 2008)



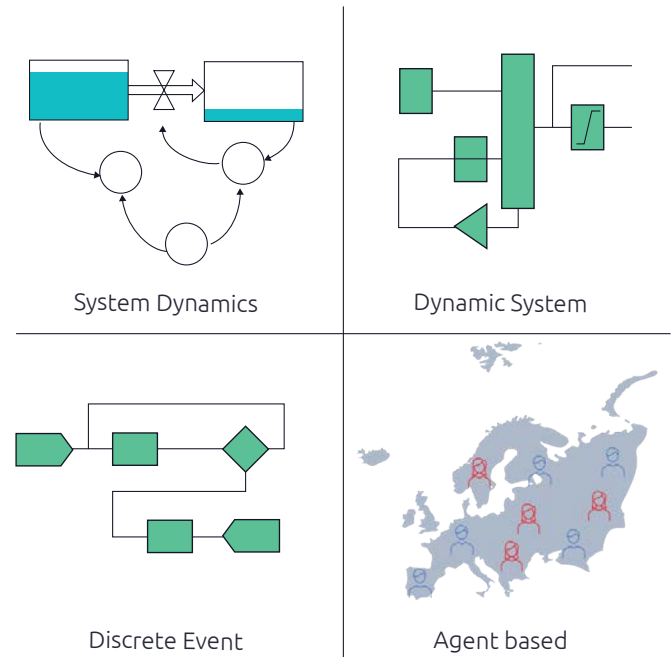
1. S. Robinson, "Conceptual modelling for simulation part i: Definition and requirements" Journal of the operational research society, vol. 59, no. 3, pp. 278–290, 2008. <https://cyberleninka.org/article/n/1101765.pdf>

METHODEN IN SIMULATIEMODELLERING

Er zijn ten minste vijf hoofdmethoden in simulatiemodellering (figuur 2):

- **Agent-Based:** Agent-Based Simulation (ABS) wordt gemodelleerd door het analyseren van agents die met elkaar interacteren. Deze simulaties worden meestal gebruikt wanneer een menselijke (agent) factor betrokken is zoals bijvoorbeeld bij luchthavens, supermarkten, en assemblagelijnen.
- **Discrete-Event:** Discrete Event Simulation (DES) is een methode om real-world systemen te simuleren door middel van discrete tijdstappen, bijvoorbeeld gebeurtenissen. Elke gebeurtenis en elk proces wordt gesimuleerd en krijgt een specifiek tijdstempel toegewezen. Kleinere tijdstappen (wat betekent dat er meer gebeurtenissen zijn) maken de simulatie nauwkeuriger, maar verhogen de berekeningstijd.
- **System dynamics:** De System Dynamics Simulation (SDS)-methode beschrijft het systeem en de invloed van de omgeving op macroniveau, gebaseerd op oorzaak-gevolg relaties, reactievertragingen en feedback loops. Het toepassingsgebied van SD omvat ook stedelijke, sociale en ecologische systemen. In SD worden de real-world processen weergegeven in termen van voorraden (bijvoorbeeld van materiaal, kennis, mensen, geld), stromen tussen deze voorraden en informatie die de waarden van de stromen bepaalt.
- **Dynamic systems:** Een methode die wordt gebruikt om complexe dynamische systemen te beschrijven met behulp van geavanceerde wiskunde om te voorspellen of iets stabiel blijft of niet. Deze systemen zijn doorgaans te vinden in de regeltechniek.
- **Hybrid modelling (ABS and DES):** Een hybride aanpak met een combinatie van ABS met DES. Dit type simulatie kan zeer krachtig zijn in situaties waar processen moeten worden gemodelleerd waarbij stochastiek betrokken is. Een voorbeeld hiervan is een assemblagelijnpductieproces waarbij onderdelen moeten worden geassembleerd tot een eindproduct. Discrete Event Simulation kan de processen modelleren en de nauwkeurigheid van de assemblagelijnpeductielijn voor elke tijdstap simuleren. De agenten kunnen worden gebruikt om onderdelen, orders en operators te modelleren, waarbij elke agent zijn eigen specifieke set vooraf gedefinieerde parameters heeft.

Figuur 2 - The family of simulation modelling (Borshchev & Filippov, 2004)²



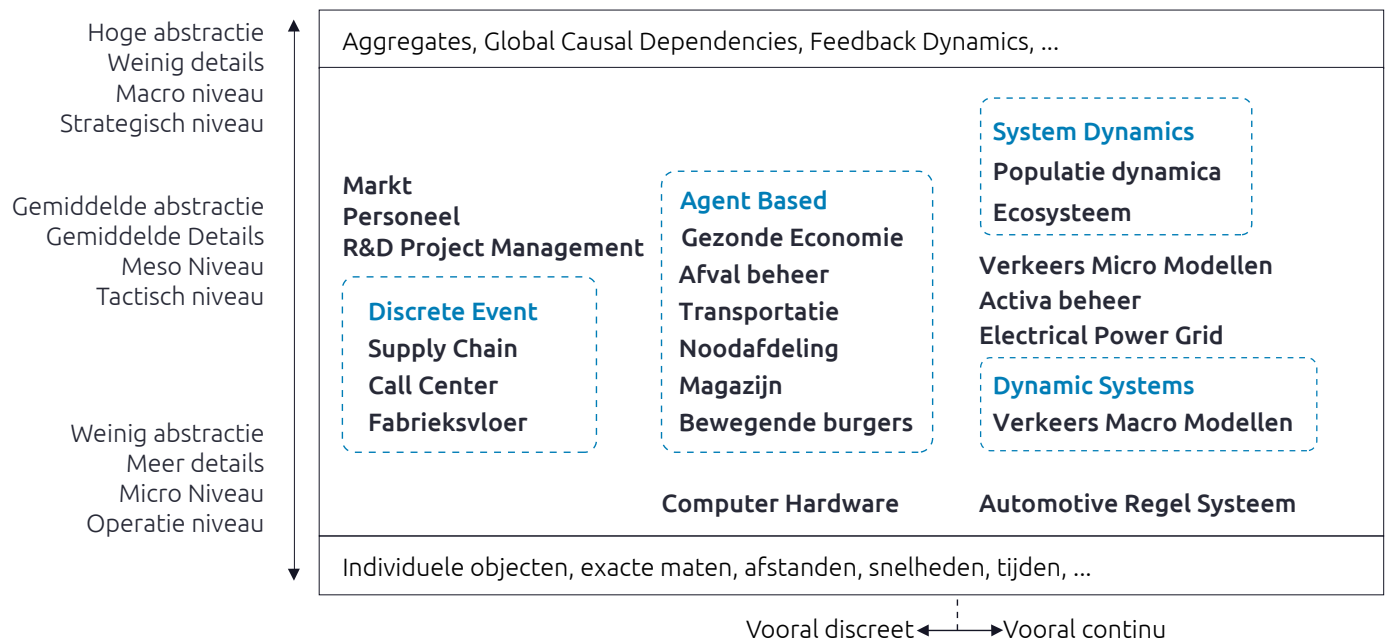
ABSTRACTIENIVEAUS EN TYPEN SIMULATIES

Figuur 3 toont de reeks problemen die kunnen worden aangepakt met simulaties. Deze problemen zijn gerangschikt op een schaal van abstractieniveau en detail. Op hogere abstractieniveaus worden problemen op strategisch niveau geanalyseerd en is het detailniveau van het model laag. Individuele elementen zoals mensen, onderdelen en producten worden op dit niveau niet beschouwd. Als we naar het micro-niveau kijken, dan zien we het tegenovergestelde. Individuele elementen worden gemodelleerd en elk detail kan van belang zijn. Simulatie van de productie valt onder het middelgrote detailniveau, omdat in plaats van elk station of elke transportband afzonderlijk te modelleren, over het algemeen gemiddelde timings gebruikt worden bij de modellering.

De simulatiemethoden die eerder zijn genoemd, worden ook in deze figuur weergegeven. Technisch gezien werken Discrete Event en Agent Based simulaties voornamelijk met discrete tijd, terwijl System Dynamics en Dynamic Systems zich bezighouden met continue processen. System Dynamics wordt toegepast waar het abstractieniveau het hoogst is. Voor fabrieksprocessen en supply chain processen is Discrete Event Simulation het meest geschikt en de Agent Based methode is geschikt van laag tot hoog abstractieniveau. Deze simulatiemethoden worden niet alleen gebruikt in de maakindustrie. Ook andere industrieën hebben soortgelijke problemen die met simulaties kunnen worden aangepakt.

2. Borshchev and A. Filippov, "From System Dynamics and Discrete Event to Practical Agent Based Modeling: Reasons, Techniques, Tools", The 22nd International Conference of the System Dynamics Society, July 25–29, 2004, Oxford, England. <http://www2.econ.iastate.edu/tesfatsi/systemdyndiscreteeventabmcompared.borshchevfilippov04.pdf>

Figuur 3 - Approaches and applications of simulation modeling (Borshchev & Filippov, 2004)

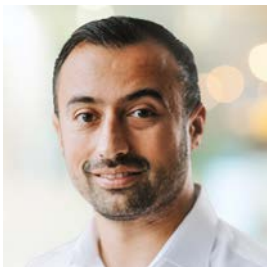


CONCLUSIE

In een wereld van complexe productieprocessen is plant simulation een X-factor. Door onze verkenning van real-world toepassingen en onze ervaringen hebben we besproken hoe plant simulation het landschap van de productie en de ondersteunende IT-processen transformeert.

Het optimaliseert de bedrijfsvoering, integreert IoT-technologieën en leidt ons naar de digital twin, de toekomst van de slimme productie. Hiermee kunnen we mogelijkheden ontsluiten zoals realtime monitoring, procesoptimalisatie op basis van echte gegevens, voorspellend onderhoud en het verbeteren van datagestuurde besluitvorming met IT-systemen voor de productie.

OVER DE AUTEUR



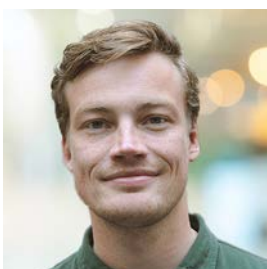
ORKHAN ALLAHVERDIYEV

Industrial Engineer

Orkhan heeft meer dan 6 jaar ervaring als industrial engineer in verschillende industrieën. Hij is consultant bij de divisie Digital Manufacturing van Capgemini Engineering. Zijn interesses liggen op het gebied van simulatie-, productie- en Industrie 4.0-projecten.

<https://www.linkedin.com/in/allahverdiyevorkhan/>

orkhan.allahverdiyev@capgemini.com



TWAN HUIJBERS

Industrial Engineer

Twan is een engineer met meer dan 5 jaar ervaring in mechanical en industrial engineering. Zijn interesses liggen op het gebied van simulatie en data binnen het vakgebied van engineering. Hij is lid van het team Hybrid Intelligence van Capgemini Engineering.

<https://www.linkedin.com/in/twan-huijbers/>

twan.huijbers@capgemini.com

13

ROBUUSTHEID VAN SOFTWARE ENGINEERING DOOR MIDDEL VAN HYPER-AUTOMATISERING

AUTOMATISERING OP HOOG NIVEAU: HYPER-AUTOMATISERING



HIGHLIGHTS

- “Robuuste engineering” is erop gericht producten te creëren die van hoge kwaliteit, betrouwbaar en kosteneffectief zijn^{1,2}.
- De snelheid van softwareontwikkeling moet omhoog.
- De behoefte aan Software engineers groeit 5-10x sneller dan de wereldbevolking.
- Groei door alleen maar meer engineers aan te nemen is een anti-patroon voor Robuustheid en Software Kwaliteit.
- Wij zien Hyper-automatisering als de belangrijkste onderscheidende factor in het software-engineeringproces en geloven dat robuustheid niet kan worden bereikt zonder hyper-automatisering.

Robuuste engineering is erop gericht producten te creëren die bestand zijn tegen variaties in hun omgeving en gebruik en welke tevens van hoge kwaliteit, duurzaam, betrouwbaar en kosteneffectief zijn.

ROBUUSTHEID

De Taguchi-methode is een populaire benadering van robuuste engineering. De methode werd na de Tweede Wereldoorlog ontwikkeld door Dr. Genichi Taguchi en wordt door veel bedrijven over de hele wereld gebruikt om de productkosten te verlagen, de kwaliteit te verbeteren en de ontwikkelingstijd te verkorten.

De snelheid van softwareontwikkeling is één van de overlevingsfactoren van bedrijven geworden om zich te onderscheiden van de concurrentie. Groeiende softwarestacks, snellere productleveringen, groeiende behoefte aan functies, en een sneller personeelsverloop van ingenieurs, stellen steeds hogere eisen aan het softwareontwikkelingsproces.

Wij zien hyper-automatisering als de belangrijkste onderscheidende factor in het software-engineeringproces. Het maakt een strategie mogelijk om fouten zo vroeg mogelijk in het ontwikkelproces te voorkomen, door het gebruik van formele modelleringstalen, het genereren van code en geautomatiseerde verificatie en validatie. Wij geloven dat onze klanten met hyper-automatisering de robuustheid van de softwaresystemen die zij ontwikkelen aanzienlijk kunnen verbeteren.

AUTOMATISERING

In het algemeen proberen we met automatisering een proces te versnellen en waarde te verhogen. Het zorgt voor kortere feedbackloops en minder menselijke fouten in werk met repetitieve stappen, met een hogere kwaliteit en robuustheid tot gevolg.

We kunnen automatisering zowel in de primaire waarde stroom toepassen (het bouwen van het product) als in de secundaire waarde stroom (het bouwen van de ontwikkelomgeving).

Er zijn twee algemeen gebruikte strategieën om waarde te verhogen:



- **Vergroot het aantal engineers dat waarde creëert.** Deze strategie raden we af omdat dit hooguit een oplossing is voor de korte termijn. Ook is dit een erg dure oplossing omdat er voor elke engineer kosten moeten gemaakt worden voor het rekruteren en stroomlijnen van processen, en er blijvend kosten gemaakt moeten worden voor educatie, licenties, tooling e.d.

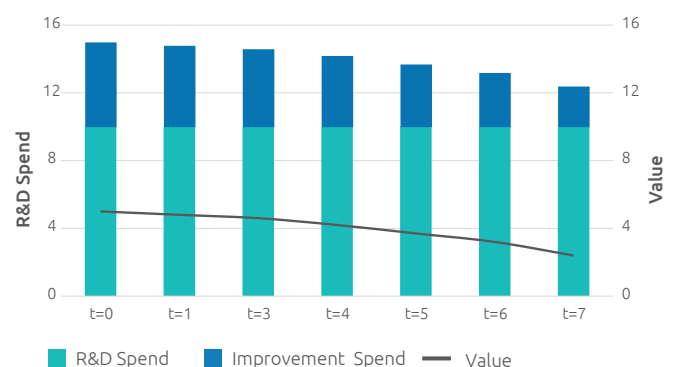


- **Vergroot de hoeveelheid waarde creatie per engineer.** Dit heeft onze voorkeur waarbij we zien dat vooral in de secundaire waarde stroom extra veel waarde toegevoegd kan worden door middel van automatisering.

Bij traditionele R&D investeringen is het doel om bij een jaarlijkse investering van X Euro daar elk jaar Y Euro aan waarde uit terug te krijgen maar meestal wordt dit niet gerealiseerd. De ROI op een R&D investering daalt meestal bij een gelijkblijvende jaarlijkse investering.

Dit patroon kunnen we doorbreken met behulp van Hyper-automatisering.

Figuur 1: R&D Spend vs ROI (reality)



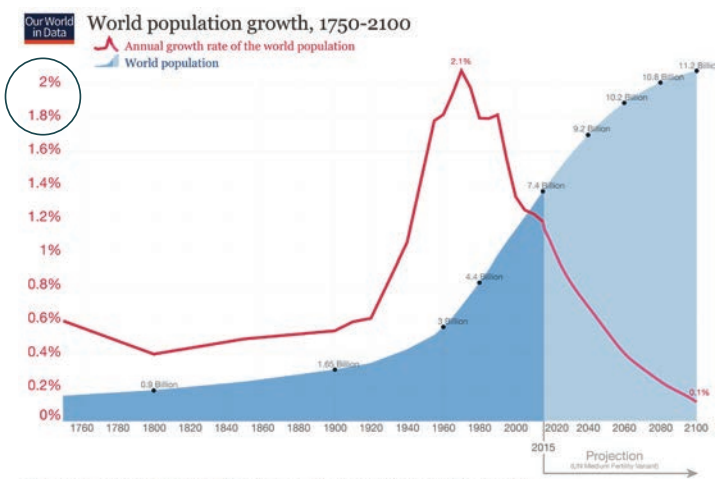
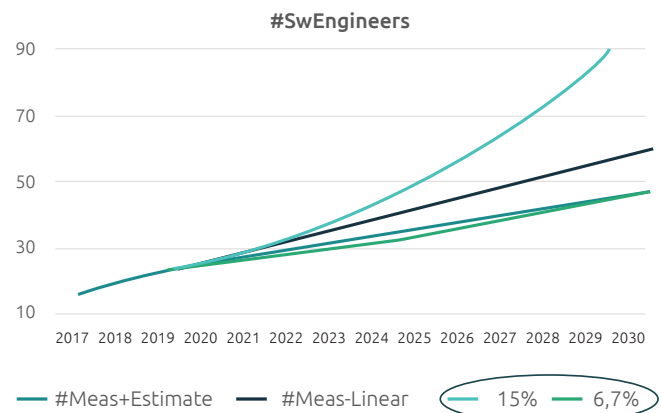
1. “Robust Engineering: Learn How to boost Quality While Reducing Costs & Time to Market” by Genichi Taguchi, Subir Chowdhury, Shin Taguchi
 2. “Robust Engineering” by Shin Taguchi

WAAROM AUTOMATISERING ALLEEN NIET GENOEG IS

Er zijn een aantal belangrijke mondiale trends die ons aanzetten om 'meer' te doen:

- **Time-to-Market**
Het te laat betreden van de markt heeft een verlies in marktpositie tot gevolg.
- **Toenemende handhaving van regelgeving**
De huidige regelgeving vereist dat de kwaliteit, duurzaamheid en robuustheid van onze producten verbeterd moeten worden. Acceptabele kwaliteit van gisteren is vandaag de dag niet langer acceptabel.
- **De Software-Explosie**
De software in producten worden steeds belangrijker en omvangrijker. Adobe Photoshop bijvoorbeeld bestond uit 10 duizend regels code in 1990 en maar liefst 4,5 miljoen regels in 2020.
- **Toenemende domein complexiteit** Voor het domein van de deeltjesfysica zijn steeds complexere systemen nodig, zoals het CERN.
- **Toename van de historische complexiteit**
Bij een systeem zoals een serverrek wordt er in de loop der jaren steeds meer aan toegevoegd, waardoor het onoverzichtelijk dreigt te worden.
- **De wereld biedt niet voldoende talent om aan de vraag naar Software engineers te voldoen**
De behoefte aan Software-engineers groeit 5-10x sneller dan de wereldbevolking³. Dit betekent dat we niet kunnen blijven schalen met het aantal software-engineers.

Figuur 2: Verwachte aantal benodigde software engineers



Figuur 3: Daxx³

3. DAXX, Feb 2020, <https://www.daxx.com/nl/blog/ontwikkeling-trends/aantal-software-ontwikkelaars-developer-statistics>



Daarnaast gaat een groei in het aantal software engineers ten koste van de robuustheid:

- Mensen maken fouten, dus een verdubbeling van het aantal mensen verdubbelt ook het aantal fouten (denk aan het begrip 'maandagochtendauto'⁴).
- We weten dat er gemiddeld per duizend regels code een X aantal fouten gemaakt worden. Verdubbeling van de code betekent dus ook een gemiddelde verdubbeling van het aantal fouten.
- Groei van een organisatie veroorzaakt groeijs op het gebied van management en van het integreren/trainen van de nieuwe engineers in de bedrijfscultuur en werkwijze, met een meetbare verlaging in snelheid en kwaliteit tot gevolg.

We kunnen hieruit concluderen dat het alleen maar groeien met het aantal software-engineers een anti-patroon is voor robuustheid en software kwaliteit.

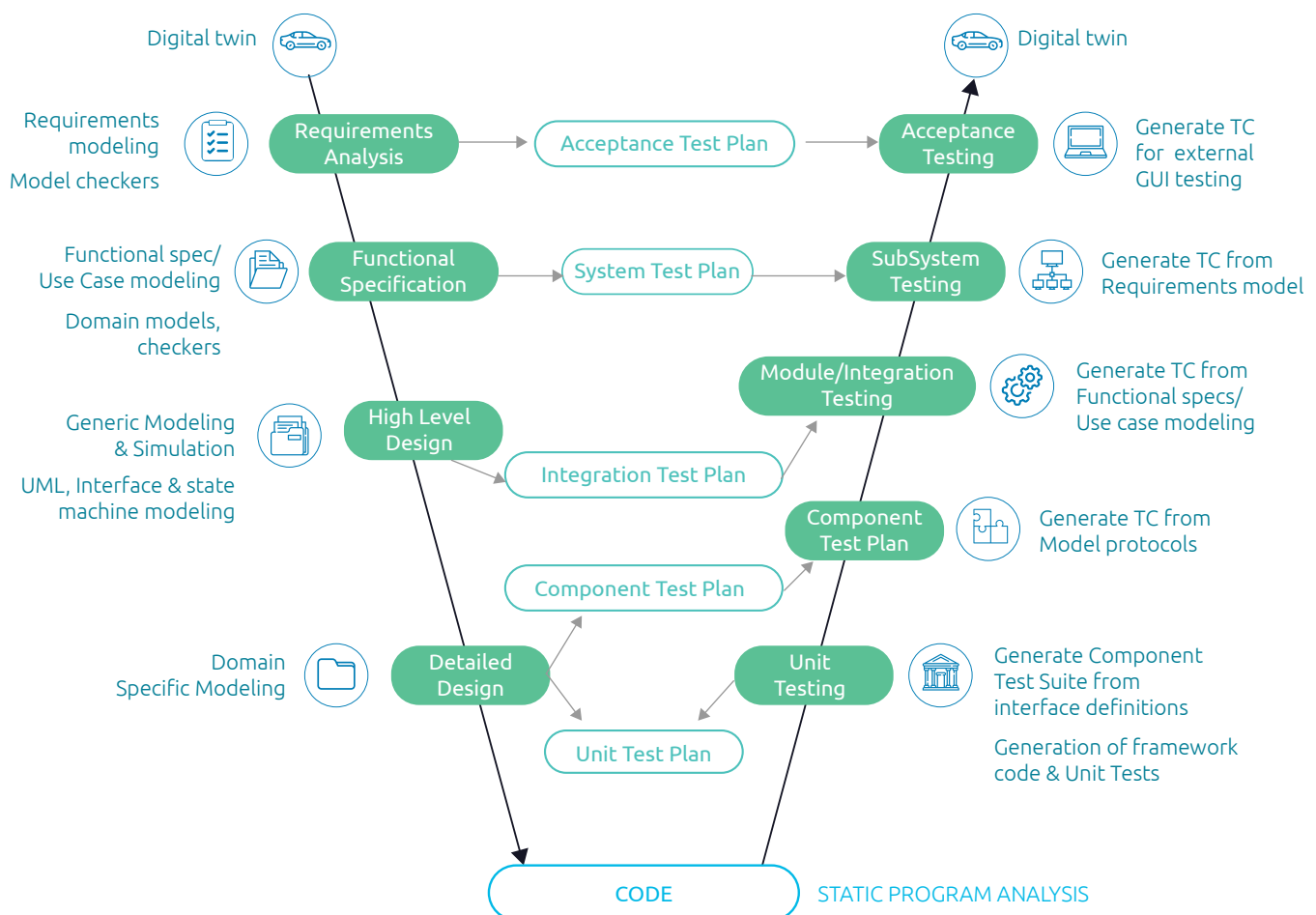
Als we naar het V-model kijken, zien we dat automatisering momenteel voornamelijk aan de rechterzijde plaatsvindt⁵

Denk aan statische code checks, test- en bouwactiviteiten die geautomatiseerd worden waarvoor Quality gates worden ingericht.

Hierbij wordt echter voorbijgegaan aan alle andere aspecten zoals de eerdergenoemde trends en de groei in kennis en de organisatie waar engineers mee te maken krijgen.

Het is belangrijk dat een bedrijf blijft focussen op hun kernactiviteiten en alle context activiteiten bij voorkeur (hyper-)automatiseert zodat de eigen engineers zich niet bezig hoeven te houden met het eigen maken van nieuwe tools, technieken en standaarden en zich kunnen concentreren op het bouwen van het product.

Figuur 4: Hyper-automatisering V-model



4. <https://www.ensie.nl/betekenis/maandagmorgenexemplaar-maandagochtendexemplaar-maandagauto> & <https://topgear.nl/autonieuws/maandagochtendmodellen-bestonden-echt>
 5. <https://en.wikipedia.org/wiki/V-model>

HYPER-AUTOMATISERING

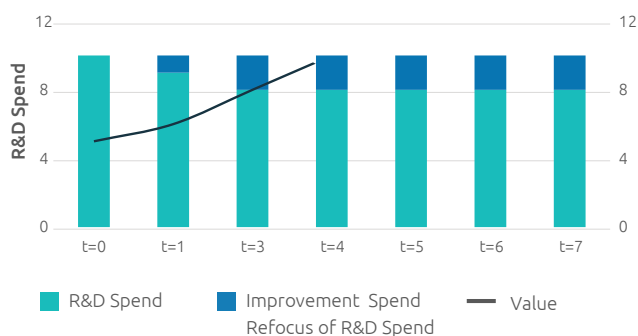
Door middel van het verschuiven van een (gedeelte) van het R&D budget naar hyper-automatisering kunnen we dit ombuigen naar een stijging van de ROI.

Met hyper-automatisering van tools en middelen voor de primaire waarde stroom, zorgt hyper-automatisering ervoor dat engineers zich kunnen focussen op hun kernactiviteiten en steeds sneller hun waarde kunnen opleveren.

Om het verschil uit te leggen tussen automatisering en hyper-automatisering nemen we het voorbeeld van een shuttlebus op een vliegveld die passagiers van en naar hun vliegtuig brengt.

Eén van de kerntaken van een vliegveld is om de passagiers van en naar het vliegtuig te brengen. Dit is een dynamisch

Figuur 5: ROI met hyper-automatisering



Figuur 6: Automatisering en hyper-automatisering

AUTOMATISERING

VERVOER VAN A NAAR B.

Vereist een contextdomeinexpert; een chauffeur.



- Chauffeur beschikt over rijcompetenties
- Verbeter de vaardigheden van de bestuurder per rit
- Past rijvaardigheden toe om intentie te vertalen naar interface-instructies, omgevingsomstandigheden en UI-feedback te interpreteren

HYPER-AUTOMATISERING

BRENG MIJ NAAR B.



- Intentie uitdrukken
- Shuttlefabrikant beschikt over rijcompetenties
- Leer van de ervaringen van alle ritten en verbeter de vaardigheden van de shuttles.

SOCIETY OF AUTOMOTIVE ENGINEERS (SAE) AUTOMATION LEVELS

00	01	02	03	04	05
No Automation Zero autonomy, the driver performs all driving tasks.	Driver Assistance Vehicle is controlled by the driver, but some driving assist features may be included in the vehicle design.	Partial Automation Vehicle has combined automated functions, like acceleration and steering, but the driver must remain engaged with the driving tasks and monitor the environment at all times	Conditional Automation Driver is a necessity, but is not required to monitor the environment. The driver must be ready to take control of the vehicle at all times with notice	High Automation The vehicle is capable of performing all driving functions under certain conditions. The driver may have the option to control the vehicle	Full Automation The vehicle is capable of performing all driving functions under all conditions. The driver may have the option to control the vehicle

Waarde verhoging door hyper-automatisering

Door te hyper-automatiseren hebben engineers meer 'tijd' om aan kernactiviteiten te werken en dus waarde toe te voegen. Menselijke fouten worden verminderd hetgeen de kwaliteit, robuustheid en Time-to-Market ten goede komen.



De traditionele aanpak om de waarde per engineer te verhogen is om steeds meer standaarden en generieke/COTS tools toe te voegen in het software-ontwikkelproces, welke software-engineers moeten leren beheersen. Dit vereist meer kennis en opleiding voor elke engineer. Niet elke engineer zal dit even gemakkelijk en snel beheersen wat vaak te herkennen is aan het feit dat er binnen de organisatie dan gefocust wordt op 'senior' engineers die in het team nodig zijn om succesvol te kunnen worden.



Bij hyper-automatisering zien we een andere aanpak: Die van een op maat gemaakte omgeving met hyper-geschaalde automatisering met focus op toegevoegde waarde. De kennis en kunde die de engineers zouden moeten opdoen voor deze nieuwe tooling, wordt nu verborgen in het geautomatiseerde deel, waardoor we ook minder afhankelijk worden van senior-level engineers.

Hyper-automatisering passen we toe met de systeem denkmethode. Niet alleen aan de rechter kant van het V-Model passen we het toe, maar juist gedurende het gehele engineering proces waarbij met name aan de linker kant van het V-Model nog veel winst valt te behalen.

Onder hyper-automatisering vallen de volgende maatregelen:

- Model-Driven (Software) Engineering ondersteund door geautomatiseerde tools.
- Interfacemodellering en het genereren van interfacecodes, bijvoorbeeld met COTS-tools zoals ASD, Dezyne, CocoTech e.d.
- Generieke modellering en codegeneratie zoals U/Sys-ML, State Machine-modellering e.d.

- Het automatiseren van alle repetitieve taken om menselijke fouten te verwijderen en de snelheid te verhogen.
- Domein specifieke modellering en codegeneratie: Hyper-automatisering om domein specifieke kennis en vaardigheden te 'verbergen'.

Al deze maatregelen hebben een effect op kwaliteit en het is belangrijk dat een organisatie niet valt voor de mythe dat er een afweging gemaakt moet worden tussen snelheid en kwaliteit. Uit empirische gegevens blijkt dat we code van hoge kwaliteit nodig hebben om snel te kunnen handelen⁶.

SAMENVATTEND

- **Hyper-automatisering**
Automatiseer alle repetitieve taken (met behulp van experts/partners) en doe dit over het gehele engineering proces.
- **"Get rid of bugs"**
Het is kosteneffectiever om bugs te voorkomen dan deze tijdens het testen op te sporen en te herstellen door bijvoorbeeld gebruik te maken van coderingsrichtlijnen, interfacemodellering en MD(S)E-initiatieven.
- **"Shift Left" & Testen**
Het is belangrijk om niet alleen een goede agile testpiramidestrategie te implementeren,^{7,8,9} maar een 'Shift Left' door te voeren naar de linkerkant van het V-Model.
- **"Fail Fast / Early Feedback"**
Het is bekend dat korte feedbackcycli de kwaliteit verbeteren. Hogere kwaliteit betekent dat we in staat zijn te versnellen.
- **Handhaving**
'Regels' werken niet als ze niet worden gehandhaafd. Definieer en meet KPI's voor het gebruik in quality gates, overall in het V-Model.

6. Capers, Jones. Applied Software Measurement: Global Analysis of Productivity and Quality. 1996

7. 2022 Paper "Code Red: The Business Impact of Code Quality – A Quantitative Study of 39 Proprietary Production Codebases"

Adam Tornhill (CodeScene) & Markus Borg (RISE Research Institutes of Sweden, Lund University Lund, Sweden). <https://arxiv.org/abs/2203.04374v1>

8. Mike Cohn, 'Test Pyramid' Succeeding with Agile, 2009

9. Agile testing: A practical guide for testers and agile teams

OVER DE AUTEUR



RICHARD VAN DE LAARSCHOT

Chief Solution Architect

Richard leidt het Center of Excellence "Quality Assurance" met als doel om sneller en tegen lagere kosten, hogere kwaliteit software te leveren door middel van het automatiseren van software-engineering en test processen.

<https://www.linkedin.com/in/richardvandelaarschot/>

richard.vandelaarschot@capgemini.com

COLOFON

HET CAPGEMINI APPLICATIONS UNLEASHED TEAM 2024

Editor in Chief: Rik Pennartz

Design: Puja Sengupta, Johanna Achterberg

Branding: Els van der Velden, Pamela Datta

Project en Marketing Management: Thomas de Klerk, Rebecca Scott

Met speciale dank aan:

Ron Tolido
Roger Wannee
Sjoukje Zaal
Gerard Peters
Han Konings
Fina Piazza
Esmee van de Giessen
Dr. Cara Antoine

Thought Leaders:

Rikkert Wijrdeman
Ron Tolido
Donald Hessing
Twan Huijbers
Orkhan Allahverdiyev
Kimberly Snoyl
Richard van de Laarschot
Malin de Silva
Sanne Visser
Menno Bruijn
Pascal Regeer
Zahir Wanga
Marc Steenbergen
Arin Roy
Philip Wagenaar
Sjoukje Zaal

OVER CAPGEMINI

Capgemini is een wereldwijde, maatschappelijk verantwoorde en multiculturele marktleider met 360,000 mensen in bijna 50 landen. Als strategisch partner ondersteunt Capgemini organisaties bij hun transformatie door gebruik te maken van de kracht van technologie. Hierbij laat de Group zich leiden door zijn bestaansreden: menselijke energie vrijmaken door middel van technologie voor een inclusieve en duurzame toekomst. Met meer dan 50 jaar ervaring en expertise in uiteenlopende sectoren, vertrouwen klanten de aanpak van hun zakelijke behoeften toe aan Capgemini: van strategie en ontwerp tot operationeel beheer. Dit gebeurt door gebruik te maken van innovaties in cloud, data, kunstmatige intelligentie, connectiviteit, software, digital engineering en platforms. De Group behaalde in 2022 een omzet van €22 miljard.

GET THE FUTURE YOU WANT | www.capgemini.nl

Capgemini Nederland B.V.
Postbus 2575 - 3500 GN Utrecht
Tel. + 31 30 203 05 00
www.capgemini.nl