Auteurs: R.E.J. (Ruben) Faber, werkt bij het NCSC als strategisch adviseur en is bereikbaar via https://nl.linkedin/in/rubenfaber en dr. ir. D.J. (Dion) Koeze, eveneens werkzaam bij het NCSC als onderzoeker, hij is bereikbaar via research@ncsc.nl.

Generatieve AI vergt meer dan code en data

In het snel evoluerende landschap van de informatietechnologie staat generatieve kunstmatige intelligentie (GenAI) aan de vooravond van een transformatie. Deze technologie, die zich richt op het creëren van content – van tekst tot beelden en zelfs code – door het leren van enorme datasets, heeft zich snel ontwikkeld tot een kerncomponent binnen de innovatiestrategieën van organisaties. Met vooruitgangen zoals OpenAI GPT-4, Google's Gemini en Anthropic's Claude-3, die mensachtige teksten kunnen genereren op basis van een brede en ongestructureerde context, is de potentie van generatieve AI onmiskenbaar.

De interesse in GenAI overstijgt industrieën en grenzen, waarbij organisaties de mogelijkheden onderzoeken of al stappen zetten richting een concrete implementatie van de nieuwe technologie. De verschillende toepassingsvormen lijken door de brede inzetbaarheid van de technologie onbegrensd en bieden Nederlandse organisaties innovatiekansen die tot voor kort ondenkbaar waren. Bijna elke organisatie, van startup tot multinational, onderzoekt momenteel de mogelijkheden. Zo kan generatieve AI bijvoorbeeld worden ingezet om de communicatie met klanten in meerdere talen te vereenvoudigen, bij het genereren en stroomlijnen van juridische en financiële documenten, om complexe logistieke processen te optimaliseren of diagnostische ondersteuning van artsen binnen de gezondheidszorgsector.

Drie pijlers

Voor een effectieve integratie van generatieve AI in bedrijfsprocessen is het cruciaal om te begrijpen dat de ontwikkeling ervan niet alleen draait om code en data, maar dat het een nieuw paradigma vereist van de toeleveringsketen. Deze zienswijze bestaat uit drie pijlers, die hun weerslag kennen in alle dimensies van IT-management, -ontwikkeling en -operatie. Pijlers worden in dit geval breed gedefinieerd als: de route door het totale procuratie- en leveringsproces; de implementatie tot aan de oplevering in een productieomgeving en integratie in bestaande systemen. Binnen de IT-toeleveringsketen is deze driedeling van essentieel belang om GenAI op een toekomstbestendige manier aan te besteden, te ontwikkelen en te implementeren.

- **Code:** programmeren, ontwikkelen, testen, implementeren, kopen of adopteren van ICT oplossingen vallen onder deze pijler. Het zijn alle activiteiten die een ICT-organisatie vanuit haar bestaansrecht kent en doet – het ontwikkelen van een applicatie die een bepaald proces ondersteunt of in een bepaalde behoefte voorziet. Het gaat hier nadrukkelijk om de brede definitie, waaronder functioneel ontwerpen, testen, reviewen – deze processen kennen traditioneel gezien hun weerslag in de code en andere artefacten die uiteindelijk in een softwareoplossing bij elkaar worden gebracht.
- **Data:** de succesvolle werking van GenAI berust op de beschikbaarheid van voldoende data van goede kwaliteit. Zowel in de ontwikkeling van foundational models, de onderliggende modellen van bijvoorbeeld OpenAI, als bij het toespitsen (finetuning) van modellen voor de toepassing in de eigen organisatie. We kennen deze pijler ook uit toepassingen die voortkomen uit de Data Science en omvat alle

aspecten van de Data-pipeline: data-acquisitie, -verwerking en -beheer.
- **Machine Learning (ML):** deze pijler is de essentiële verrijking van de traditionele tweedeling tussen code en data. Het omvat trainen, valideren en implementeren van een AI-model. Ondanks dat Machine Learning gedaan wordt via een code en op data is het van belang om dit als separate pijler te onderkennen. Een getraind model wordt in moderne toepassingen niet meer losstaand gebruikt, maar bijvoorbeeld in een pipeline met *Retrieval Augmented Generation*, een techniek waarmee bestaande informatie van een systeem ontsloten kan worden door GenAI.

Het begrijpen en beheersen van deze drie pijlers is essentieel voor het veilig en effectief implementeren van GenAI. Dit vraagt om een verschuiving in denkwijze van traditionele ICT-benadering naar een meer geïntegreerde visie die de complexiteit en de onderlinge verbondenheid van code, data en ML erkent. Door deze benadering kunnen organisaties niet alleen de vruchten plukken van GenAI, maar ook zorgen voor robuuste cybersecurity en ethische inzet van deze krachtige technologie. De transitie naar dit nieuwe paradigma vereist leiderschap dat vooruitziet, bereid is om traditionele denkpatronen te doorbreken en de waarde van een multidisciplinaire benadering onderkent. Het succesvol navigeren door deze verandering zal niet alleen technische expertise vereisen, maar ook een diepgaand begrip van de strategische, ethische en maatschappelijke implicaties van GenAI. Zo kunnen we de belofte van deze revolutionaire technologie waarmaken, terwijl we de veiligheid en integriteit van ons digitale ecosysteem waarborgen.

Fundamenten van GenAI en cybersecurity

De samenvloeiing van GenAI en cybersecurity vormt een boeiend doch complex domein dat unieke uitdagingen en kansen biedt. Terwijl generatieve AI de grenzen van technologische innovatie verlegt, brengt het ook nieuwe risico's en kwetsbaarheden met zich mee die speciale aandacht vereisen binnen het cybersecuritylandschap. Het feit dat er een apart raamwerk voor aanvalstechnieken wordt bijgehouden door MITRE – het ATLAS raamwerk, naast het bekende ATT&CK raamwerk – geeft dit ook aan. Wat maakt cybersecurity bijzonder in de context van Generatieve AI? Het aanvalsoppervlak van applicaties die gebruik maken van GenAI kent nieuwe karakteristieken. Dit kunnen we beter begrijpen door een aantal typen aanvallen te bekijken.

- **Data Poisoning en Manipulatie:** bij GenAI is de integriteit van de trainingsdata cruciaal. Een aanval waarbij de data gemanipuleerd wordt (data poisoning) kan leiden tot het genereren van valse, misleidende, of ongewenste output. Dit vraagt om geavanceerde verificatie- en validatiemechanismen om de integriteit van data te waarborgen
- **Modeldiefstal en Reverse Engineering:** GenAI modellen vertegenwoordigen aanzienlijke intellectuele en financiële waarde. Aanvallers kunnen proberen deze modellen te stelen of via reverse engineering te dupliceren. Bescherming tegen dergelijke aanvallen vereist geavanceerde technieken, zoals modelversleuteling en -watermerking.
- **Injection Attacks:** ondanks dat injectieaanvallen – waarbij malafide code of instructies ingeschoten kunnen worden – ook in webapplicaties in het algemeen nog steeds een bekende aanvalstechniek zijn volgens de OWASP Top 10, zien deze aanvallen er fundamenteel anders uit wanneer met menselijke taal of andere media zoals foto's met een applicatie te interacteren is.

Om een adequate vertaalslag te maken naar de realiteit van het aanbesteden en implementeren van ICT-oplossingen kan de driedeling in de verschillende pijlers worden gebruikt.
De **codepijler** omvat de ontwikkeling van algoritmes en software die de ruggengraat vormen van AI-systemen. Alle beveiligingsrisico's die van toepassing zijn bij het ontwikkelen van veilige softwarecode, zijn daarom ook van toepassing wanneer binnen het domein van GenAI wordt ontwikkeld: secure coding, (geautomatiseerde) code review en het adequaat beveiligen van omgevingen waar de code in wordt ontwikkeld en gedistribueerd.

Binnen de **datapijler** is de grootste zorg de bescherming van de data die wordt gebruikt voor het trainen en valideren van AI-modellen. Dit omvat maatregelen tegen ongeautoriseerde toegang, datalekken, en de ongewilde manipulatie van trainingsdata. Belangrijk hierbij is het besef dat data een steeds grotere rol gaat spelen in de toeleveringsketen, doordat GenAI rust op veel data van goede kwaliteit. Net zoals code gecompromitteerd kan worden stroomopwaarts in de keten, kan datzelfde gebeuren met de onderliggende data van modellen die worden ingezet. Verder spelen op dit domein veel bestuurlijke en juridische overwegingen, zoals de herkomst van data en bescherming van persoonlijke gegevens. Deze vraagstukken kunnen gedeeltelijk technisch worden opgelost (pseudonimi-

Het begrijpen en beheersen van de pijlers code, data en Machine Learning is essentieel voor het veilig en effectief implementeren van GenAI

sering, anonimisering), maar vergen vooral een sterk gefundeerde visie die onderdeel is van de bedrijfscultuur: wat zien wij als veilig, rechtmatig en ethisch verantwoord gebruik van data?

In de **Machine Learning-pijler** gaat het om bescherming en doorlopende bijwerking van de modelarchitectuur, toetsing en validatie van uitkomsten en onbedoelde effecten, in kaart houden van modelspecifieke kwetsbaarheden, en ketenintegriteit van het model van training tot ingebruikname. Steeds meer zal het gebeuren dat het trainen van modellen niet meer gebeurt bij een organisatie zelf of een directe leverancier, maar dat er door een keten heen modellen worden aangepast, via finetuning of few-shot learning, voordat het tot ingebruikname komt.

Concrete productontwikkeling met Generatieve AI

Om de verdeling over de drie domeinen verder in te kleuren ontwikkelen we als voorbeeldcasus een chatbot die namens een organisatie communiceert met klanten of medewerkers. Dit is nog een betrekkelijk 'eenvoudige' toepassing in vergelijking met eerder genoemde kansen, maar daarmee een oplossing waar veel organisaties momenteel concreet mee bezig zijn. Daarnaast heeft elke consument (wisselende) ervaringen met dergelijke bots. Wanneer een organisatie een chatbot wil ontwikkelen kan de driedeling in pijlers wederom worden toegepast.

Code

Bij de ontwikkeling van de chatbot is het eerste aandachtspunt de codepijler. Dit behelst het ontwerp en de implementatie van de software die de basis vormt van de chatbot. Secure-by-design principes zijn hier cruciaal: vanaf het begin moet de chatbot worden ontwikkeld met een sterke focus op de beveiliging en privacy. Dit betekent dat ontwikkelaars rekening houden met potentiële beveiligingsrisico's en kwetsbaarheden, zoals cross-site scripting of fouten in access control, en deze proactief aanpakken door middel van onder andere moderne coderingsstandaarden, code reviews, voldoende monitoring en logging, en geautomatiseerde beveiligingstests. Wanneer de chatbot ook acties kan ondernemen naast het genereren van tekst of andere media, zogenaamd function calling, om bijvoorbeeld meer gegevens op te kunnen halen of om acties uit naam van de gebruiker uit te voeren, is het belangrijker dan ooit om zorgvuldige maatregelen te nemen. Daarnaast moet de chatbot worden geprogrammeerd om te voldoen aan de GDPR en andere relevante privacywetgeving, bijvoorbeeld door gebruikers duidelijk te informeren over het gebruik van hun gegevens en hen controle te geven over hun persoonlijke informatie.

Data

De datapijler betreft het verzamelen, verwerken en beheren van de data die de chatbot gebruikt om te leren en te functioneren. Dit omvat zowel de initiële trainingsdata, de data die gebruikt wordt bij fine-tuning of few-shot learning, alsook de voortdurende input van gebruikersinteracties. Het waarborgen van de kwaliteit en diversiteit van deze data is essentieel voor de effectiviteit van de chatbot. Om bias en onnauwkeurigheden te voorkomen, moeten data scientists en ontwikkelaars zorgvuldig selecteren welke data wordt gebruikt voor het trainen van de chatbot. Dit begint bij de selectie van het model waarop de applicatie wordt gebaseerd en loopt door tot de laatste prompt engineering die nodig is om de bot de gewenste interactie te geven. Tegelijkertijd moeten ze strikte privacyrichtlijnen volgen om te zorgen dat persoonsgegevens beschermd worden. Dat betekent dat alle verzamelde data geanonimiseerd of gepseudonimiseerd moet worden en dat data alleen voor specifieke, gerechtvaardigde doeleinden wordt gebruikt.

Machine Learning

Ten slotte omvat de Machine Learning-pijler het trainen, valideren en implementeren van de AI-modellen die de chatbot in staat stellen om te leren van interacties en in de loop van de tijd te verbeteren. Dit proces moet zorgvuldig worden beheerd om te

zorgen voor de betrouwbaarheid en veiligheid van de chatbot. Dit houdt in dat er maatregelen worden getroffen om overfitting te voorkomen, dat er validatiestappen worden ingebouwd om de accuraatheid van de chatbot te verzekeren en dat er voortdurend wordt gemonitord op mogelijke veiligheidsrisico's die kunnen ontstaan door manipulatie van de input, bijvoorbeeld via adversarial attacks. Daarnaast is het van belang dat het model regelmatig wordt geëvalueerd en bijgewerkt op basis van nieuwe data en feedback van gebruikers, om zo te zorgen voor een continue verbetering van de prestaties en gebruikerservaring.

Door de drie pijlers zorgvuldig toe te passen en te integreren in het ontwikkeltraject van een chatbot, kan een organisatie een krachtige tool ontwikkelen die niet alleen effectief communiceert met klanten, maar dit ook doet op een veilige, betrouwbare en ethisch verantwoorde manier. Het succes van dit traject hangt af van een multidisciplinaire aanpak waarbij ontwikkelaars, datawetenschappers, cybersecurity experts, en juridisch adviseurs samenwerken om de uitdagingen en mogelijkheden van elke pijler te adresseren.

De driedeling effectief implementeren

In het hart van de moderne digitale transformatie ligt een fundamentele verschuiving in de manier waarop we denken over en werken met technologie. Deze verschuiving, gedreven door de opkomst van GenAI en de immer aanwezige noodzaak voor robuuste cybersecurity, vraagt om een herziening van traditionele ICT-strategieën. Het is een uitnodiging aan managers en bestuurders om niet alleen technologische vernieuwers maar ook culturele architecten binnen hun organisatie te zijn. Het implementeren van de driedeling in pijlers vormt de kern van deze transformatie en stelt medewerkers in staat om op een andere manier te kijken naar de kansen en risico's van GenAI. Hoe kunnen leiders deze transitie zo effectief mogelijk begeleiden en een cultuur creëren die deze nieuwe benadering omarmt?
Het begint bij de erkenning dat de integratie van generatieve AI en cybersecurity verder gaat dan technologie alleen: het raakt aan de wijze waarop teams samenwerken, hoe projecten worden geleid en hoe succes wordt gemeten. Leiders moeten allereerst de unieke waarden en vereisten van elke pijler begrijpen en vaststellen hoe deze bijdraagt aan het grotere geheel.

Bij de codepijler is het van belang dat ontwikkelaars niet alleen schrijven wat functioneel, maar ook wat veilig en veerkrachtig is. Dit vereist een verschuiving naar secure-by-design principes, waarbij veiligheid vanaf het begin onderdeel is van de ontwikkelingscyclus. De datapijler vereist een cultuur die de waarde van data erkent en respecteert. Dit betekent dat teams werkwijzen en vaardigheden ontwikkelen op het vlak van het ethisch verzamelen, verwerken en gebruiken van data, met een duidelijk begrip van privacy- en beveiligingsrisico's. De Machine Learning-pijler introduceert complexiteit rondom het trainen en implementeren van modellen die zowel effectief als veilig zijn. Dit vereist nauwe samenwerking tussen datawetenschappers, cybersecurity experts en de rest van de IT-organisatie, om te zorgen dat modellen niet alleen nauwkeurig zijn, maar ook bestand tegen manipulatie en misbruik.

Een effectieve implementatie van deze driedeling vraagt om een cultuuromslag binnen de organisatie. Dit betekent dat leiderschap niet alleen moet komen vanuit het management, maar ook vanuit een breed gedragen visie en enthousiasme voor de nieuwe benadering van technologie en beveiliging volgens de voorgestelde driedeling.

Een dergelijke cultuuromslag begint met onderwijs en bewustwording. Bestuurders en managers moeten zorgen voor regelmatige trainingen en workshops die niet alleen de technische aspecten van AI en Cybersecurity behandelen, maar ook de ethische en strategische implicaties ervan. Het is essentieel dat alle medewerkers, ongeacht hun functie, begrijpen hoe hun werk aansluit op de bredere digitale strategie van de organisatie.
Organisaties moeten daarnaast investeren in hun capaciteiten, zowel in eigen specialistisch ML-personeel als het aan- of uitbesteden van ML-activiteiten en -componenten. Een geïntegreerde personeels- en leveranciersstrategie, die past in een visie gebaseerd op de drie pijlers, zal helpen om de juiste nieuwe collega's te vinden en ML-leveranciers te selecteren.
Als laatste is het belangrijk om een omgeving te creëren waarin experimenten, falen en leren worden aangemoedigd. Innovatie komt niet voort uit het strikt volgen van de regels, maar uit het durven verkennen van nieuwe ideeën en het accepteren dat niet elke poging succesvol zal zijn. Dit vereist een leiderschapsstijl die autonomie ondersteunt, initiatief aanmoedigt en waardeert, en de nadruk legt op continu verbeteren. Door het stimuleren van deze vrijheden, binnen de context van de drie pijlers en hun vereisten, is het mogelijk om de enorme snelheid van het AI-domein bij te houden en daar, op een cyberveilige manier, de vruchten van te plukken. Door het stimuleren van deze vrijheden, binnen de context van de drie pijlers en hun vereisten, is het mogelijk om de enorme snelheid van het AI-domein bij te houden en daar, op een cyberveilige manier, de vruchten van te plukken.