Geïnterviewde: Dasha Simons is Managing Consultant voor Trustworthy AI bij IBM Consulting en adviseert klanten over eerlijkere en transparantere AI-ontwikkeling. Ook is zij onderdeel van Team Europe Direct met de focus op Trustworthy AI bij de Europese Commissie. Zij studeerde als 'best graduate' af aan de Technische Universiteit in Delft (industrieel ontwerpen 2019). Dasha is bereikbaar via: https://www.linkedin.com/in/dasha-simons.

Op zoek naar een eerlijk AI-model

Wat is een eerlijk AI-model? Hoe beperk je risico's op desinformatie? De belangrijkste uitdaging ligt in het bepalen wat in elke specifieke toepassing als eerlijk wordt beschouwd en wat de meest geschikte manier is om de werking van het systeem uit te leggen. IBM introduceerde watsonx. Een AI en dataplatform met een set van AI-assistenten die bedrijven helpen om de impact van AI te schalen en te versnellen met betrouwbare data.

Over dit thema hebben wij met Dasha Simons van IBM gesproken en haar gevraagd naar haar inzichten, ervaringen en lessen. Onderstaand het verslag van dat interview.

Ik kwam bij IBM terecht door mijn afstudeerproject en ontdekte dat IBM een van de weinige, grote techbedrijven is bij wie het hen in het businessmodel niet gaat om de verkoop van data. IBM omarmde al vroeg de ethische principes en was daarmee één van de eerste die daarop inzette. De reden van mijn keuze voor het vak was tweeledig: aan de ene kant, bijna filosofisch nadenken over de vraagstukken waarmee je bezig bent. Jezelf vragen stellen als: wat vinden wij eerlijk? Wat omvat daadwerkelijke transparantie? Wat is voldoende transparantie zonder schending van het privacyrecht? Aan de andere kant betekent dit binnen het vakgebied van Kunstmatige Intelligentie het vertalen naar technische oplossingen die ervoor moeten zorgen dat keuzes op filosofisch, sociaal en politiek gebied hun weerslag vinden in de technische realiteit van alledag. Dat is een interessante en intellectuele uitdaging. Daarnaast draag ik als millennial bij aan een iets betere wereld – al is het een heel klein beetje.

Een kijkje bij IBM en de AI-wereld

Er zijn verschillende vormen van AI en de meest recente is die van generatieve AI (GenAI). De meer traditionele vormen zoals Machine Learning bestaan al langer, maar het grote publiek is pas recentelijk met de lancering van ChatGPT met GenAI in aanraking gekomen. Feitelijk bestaan er al GenAI modellen vanaf circa 2012. GenAI is gebaseerd op onder andere foundation models. Een voorbeeld van een Machine Learning-model is een model dat kredietrisicovoorspellingen kan doen. Ook daarvoor moet je veel data verzamelen en het model trainen. Het model beperkt zich enkel tot die taak en is zeer specifiek.

Het verschil met de Foundation Models is dat je met één model traint. Dat vereist heel veel data en veel rekenkracht. Dat model train je niet met slechts één doel voor ogen. Neem de Large Language Modellen (LLMs), de grote taalmodellen die gebruikmaken van foundational models. LLMs zijn heel goed in het voorspellen van het eerstvolgende woord in een tekst. Zo'n model is goed toepasbaar op veel laag risico taken en persoonlijke vragen uit de praktijk. Ze kunnen een liefdesbrief of gedicht schrijven of feedback geven op jouw CV, zonder of met beperkte training. Maatwerktraining is voor zoiets veel minder nodig. Dit betekent dat je het model één keer traint en op vele taken kunt toepassen. Deze modellen zijn handig voor generieke en laag risico taken. Maar bij toepassing binnen het bedrijfsleven is het wel verstandig ze toe te spitsen op de specifieke taken waarnaar je kijkt, en het extra in acht nemen van de veiligheid, de betrouwbaarheid en andere AI- en ethiekfacetten.

AI-systemen zijn uiteindelijk het verlengde van besluitvorming

Oorzaken van risico's

Het is interessant om te kijken naar de oorzaken van de risico's. Het kan helpen om te kijken naar risico's bij de input, de output en de governance. Als je naar de input kijkt, let je op vragen als: met welke data train je? En: welke risico's zijn eraan gerelateerd? Een evident voorbeeld is duidelijk te zien bij het genereren van afbeeldingen, zoals bij de Bloombergstudies (1), waarin geconstateerd wordt dat het gezicht van de gemiddelde advocaat of CEO relatief blanker en mannelijker is, terwijl het gezicht van de schoonmaakhulp relatief donkerder en vrouwelijker is. Het risico op vooringenomenheid wordt versterkt in de uitkomsten van het model. Bij tekst gebeurt hetzelfde, maar dat kan minder gemakkelijk te herkennen zijn. Dit komt doordat de trainingsdata deze vooringenomenheid bezat.

Als je naar de output kijkt, kun je ook kijken naar hoe deze misbruikt kan worden. Wat vooral te maken heeft met manipulatie en misbruik die leiden tot des(mis)informatie. Stel dat een financieel planner een simulatiemodel gebruikt om de kans te berekenen op het behalen van verschillende beleggingsdoelen in verschillende marktomstandigheden. Het model geeft echter een vertekend beeld van de werkelijkheid, doordat de uitvoerresultaten significant afwijken en positiever zijn dan de werkelijke uitkomsten. Dit laat een risico zien van de uitkomst van een model, dat leidt tot onjuiste conclusies en incorrecte besluitvorming.

Als we kijken naar governance is een voorbeeld van een risico het energieverbruik van generatieve modellen. Om één afbeelding te genereren is evenveel energie nodig als bij het opladen van een mobiele telefoon. Dat lijkt weinig, maar om de afbeelding tevredenstellend te genereren kan het je wel twintig tot dertig pogingen kosten. Het is belangrijk om je hiervan bewust te zijn zodat je daarop kunt mitigeren.

Risico op desinformatie

De risico's van GenAI zijn afhankelijk van het model dat je hebt en waar je het voor gebruikt. Sommige risico's zijn urgenter dan andere. Daarbij spelen onder andere de urgentie van de taak, de risicomanagementpraktijken, transparantie in datamanagement en robuuste modeltrainingprocedures een rol in. Stel dat je een model alleen intern gebruikt, door eigen medewerkers die vragen stellen met betrekking tot HR, dan is desinformatie minder waarschijnlijk. Maar gebruik je een ChatGPT of een andere vraag-consumentenbot voor het algemene publiek, dan kan iedereen een vraag formuleren die uiteindelijk qua output schadelijk zou kunnen zijn. Bijvoorbeeld, als je om een recept had gevraagd voor je avondeten, zou dit kunnen resulteren in het verkrijgen van recepten die uiteindelijk giftige stoffen bevatten. Een gebruiker of iemand die schade wil toebrengen aan het product, kan op tal van manieren proberen te achterhalen hoe hij het systeem kan omzeilen. Voor een intern systeem is dat risico lager, er is minder risico op desinformatie.

Watsonx platform van IBM

IBM presenteert het watsonx platform met: watsonx.ai, watsonx.data en watsonx.governance (2). Hoe kan je de samenhang daarin zien? IBM watsonx is een AI en dataplatform met een set van AI-assistenten die bedrijven helpen om de impact van AI te schalen en te versnellen met betrouwbare data. Je kunt alle drie platformen: watsonx.governance, watsonx.ai or watsonx.data, separaat of samen gebruiken voor een toolkit voor model maatwerk, governance en dataopslag.
Wat watsonx.governance interessant maakt is dat het je helpt te kunnen voldoen aan regels zoals de komende EU wet- en regelgeving, ook kun je met watsonx.governance je AI governance inregelen en tegelijkertijd de manuele taken voor de data scienceteams beperken voor bedrijven.
De aankomende EU AI Act hanteert een risicogebaseerde


NEGF
BANKCALL
GOMARKF
FIXLOC
INTPRET
GETMKS
CDULOGIC
SCAXIS

Er zijn nieuwe vaardigheden en kennis nodig bij zowel management als ontwikkelteams, die juridische, filosofisch-ethische en technische kennis omvatten

aanpak waarbij niet de technologie zelf wordt gereguleerd, maar de toepassing van de technologie. Een voorbeeld daarvan is het aanbieden van essentiële publieke of private diensten, die dan kunnen vallen onder de hoog risico categorie. Je moet dan voldoen aan modeldocumentatie en een post-market monitoringsysteem bezitten. Dit om te monitoren of er sprake kan zijn van vooringenomenheid en je dat moet kunnen uitleggen. Manueel gaat je dat veel tijd kosten om al die model documentatie bij alsook up-to-date te houden.

De watsonx.governance tooling helpt je in de automatisering en het toegankelijk maken van de relevante data voor je compliance publiek alsook voor de ontwikkelaars. Ook helpt de tooling met het automatiseren van de modeldocumentatie en het standaardiseren ervan. De tooling ondersteunt het monitoren van modellen tijdens ontwikkeling en productie op bijvoorbeeld vooringenomenheid, en zorgt dat de verantwoordelijkheden duidelijk en transparant belegd zijn binnen de organisatie.

Voor bedrijven blijft het vraagstuk: 'Wat vinden wij een eerlijk model?' Daarvoor moeten goede keuzes gemaakt worden. Hoe processen moeten worden ingericht, die niet alleen door technologie kunnen worden opgelost. Dat zijn strategische keuzes op C-niveau, die moeten bepalen welke rol ze willen spelen binnen het AI-landschap: voldoen aan minimale wet- en regelgeving of ook nog proactief handelen naar andere waardes?

De nodige uitdagingen

Toen ik vijf jaar geleden begon met dit werk was het onderwerp van betrouwbaardere AI-ontwikkeling meer een onderwerp van gesprek voor bij de vrijdagmiddagborrel. Mensen vinden het een leuk onderwerp om het er over te hebben, maar niet bij een maandagochtend budget alloceringsafspraak. En dan hebben wij het over gesprekken bij klanten. Eindelijk kreeg het onderwerp in 2023 meer prioriteit bij het grote publiek. Want AI-systemen zijn uiteindelijk het verlengde van besluitvorming, want het is het eerste systeem dat dit doet. Daarmee vormen zij het verlengde van elk bedrijf in de digitale wereld en dat met impact op de reële wereld!

De grote uitdaging zit in het definiëren wat per toepassing eerlijker is of wat de juiste manier van uitlegbaarheid voor een bepaalde toepassing blijkt te zijn. Uiteindelijk zijn het politieke vraagstukken. Die zijn lastig omdat bedrijven en technici niet gewoon zijn dat soort vraagstukken op te pakken. Er zijn nieuwe vaardigheden en kennis nodig bij zowel management als ontwikkelteams, die juridische, filosofisch-ethische en technische kennis omvatten. Dat samenspel is nieuw en in beweging. Gelukkig zijn er vele bestaande methoden of ontwikkelingen om deze risico's te mitigeren, al mag daar naar mijn mening meer concrete aandacht naar toe gaan en niet alleen door evenementen te organiseren. Ik verwacht dat (Gen)AI een met het onderwerp duurzaamheid vergelijkbaar traject ingaat. Dus geen 'ethics washing' voor (Gen)AI, maar daadwerkelijk uitvoeren. Doe je mee?

Referenties

(1) https://www.bloomberg.com/graphics/2023-generative-ai-bias

(2) watsonx.ai: training, validering, afstemming, uitrol grondbeginselen en ML-modellen; watsonx.data: schaalbaarheid van AI-werklast met betrekking tot data waar dan ook opgeslagen en watsonx.governance: verantwoordelijke versnelling, transparantie en verklaarbaarheid data en AI werkstromen