Alle vragen die gesteld werden na afloop van ons webinar “Inzicht in de kwaliteit van AI” – 25 maart 2021 en de antwoorden hierop.

Q : Leunt het testen van AI meer op specialisme/focus vanuit Development of meer op traditionele test-discipline?

A : *Jeroen: het is lastig om één van beide disciplines aan te wijzen als ultimo verantwoordelijk voor AI. Testen is m.n. verantwoordelijk voor het opstellen van de juiste testcases om inzicht te verschaffen in de kwaliteit van de AI-toepassing, alsook voor het monitoren van die kwaliteit over de verschillende AI-releases. Het bepalen hoe we deze kwaliteit ‘meten’ en meetbaar maken middels metrics is weer een gezamenlijke verantwoordelijkheid samen met de developers en business stakeholders.*

Q : Controle over AI. Is dat geen onrealistisch doel gezien de exponentiële ontwikkeling hierin voor de toekomst? Hoe zorgen we ervoor dat we nooit de doemscenario’s (door meerdere guru’s in IT en filosofie geuit) tegemoet gaan. Wat is onze rol vanuit Test hierin?

A : *Robin: interessante en filosofische vraag. Als je “controle over AI” letterlijk neemt als zijnde “alles begrijpen en van tevoren goedkeuren”, is dit onrealistisch en haalt ook juist het voordeel van AI weg in termen als efficiëntie, kostenbesparing en verhoogde gebruikerstevredenheid. Daarnaast is de rol van testers niet gericht op het ethisch en filosofisch juist gebruik van software (e.g. dit kunnen we niet voorkomen).*
De huidige AI-toepassingen zijn met name gericht op specifieke toepassingen met een bepaald doel en in deze ontwikkelingen kunnen testers (zoals alle andere disciplines zoals developers, business analisten, etc.) – het liefst zo vroeg mogelijk in het proces – inzicht geven in de kwaliteit van de toepassing. Voorbeelden hiervan zijn het functioneel testen van de AI-toepassing, datasets voor het supervised / unsupervised leren van de AI-toepassing, go/no go besluit t.a.v. andere AI-versies (zowel code, alsook configuratie). Dit is een gezamenlijke verantwoordelijkheid van het gehele team (incl. business stakeholders), waar testers uiteraard een belangrijke functie in hebben voor het geven van inzicht van en adviseren over de mate van kwaliteit van de AI-toepassing.

Q : Hoe bepaal je testdoelen?

A : *Jeroen: Net als bij andere applicaties. Echter moet hier nog meer rekening gehouden worden met veranderend inzicht. Doordat de AI leert en anders gaat reageren, zullen misschien ook de doelen van de applicatie en daarmee ook de testdoelen wijzigen. We hebben ons hier vooral op functioneel testen gericht, maar non-functionals als performance en security moeten natuurlijk ook meegenomen worden.*

Q : **Wie zou alle metrics moeten invoeren (rollen)?**

A : *Jeroen: Het bepalen van waarop gemeten gaat worden in hoeverre een test aan het verwachte resultaat voldoet, is een gecombineerde effort. Ten eerste zal samen met de business gekeken moeten worden wat er belangrijk is. Daarnaast kan ook nog met verschillende gebruikersgroepen gekeken worden naar wat zij belangrijk vinden, vaak komt hier een aanvulling op de metrieken uit. Vervolgens moet, waarschijnlijk in overleg met development, gekeken worden of en hoe deze dan daadwerkelijk gemeten kunnen worden.*

Q : **In welke fase van een development-proces stellen jullie deze AI validaties, randvoorwaarden, checks etc. op? Is dat vóór-, tijdens-, of ná afloop van de development?**

A : *Jeroen: het juiste antwoord hierop is “ja, in alle fases”. Zoals ook bij traditioneel testen is het belangrijk om zo vroeg mogelijk in het traject samen op te trekken, zodat je voor de development-fase al een gezamenlijk begrip kunt krijgen over wanneer de AI-toepassing van voldoende kwaliteit is (e.g. metrics, (test)data, requirements, use cases, ontwikkeldoelstellingen en releaseplan). Tijdens de ontwikkeling stel je de testaanpak, -gevallen en –data op en voer je vroegtijdig tests uit om de AI-toepassing kwalitatief beter te maken en zowel development als de business te toetsen of er nog een gezamenlijk beeld is. Na afloop is de rol van de tester waarschijnlijk om de volgende versie weer voor te bereiden en deze te vergelijken met eerdere versies.*

Q : **Medisch komt het testen van AI misschien neer op correlaties en causaliteit. Hebben jullie daarnaar gekeken? Correlaties tussen patiëntengroepen en medicijngebruik kunnen vastgesteld worden, maar causaliteit aantonen is nog wat anders…**

A : *Jeroen: vanuit Polteq hebben we een project gedaan i.s.m. een medische entiteit op het gebied van van het herkennen van huidafwijkingen en hier een diagnose over te stellen. Hierbij hebben we uiteraard ook gekeken naar begrippen als correlatie en causaliteit. In de medische toepassingen zijn de betrouwbaarheidsintervallen nog nauwkeuriger, gezien de impact van de uitkomsten. We hebben dan ook in nauw contact met artsen gestaan om alle resultaten ook door artsen te laten beoordelen. In de praktijk zijn correlaties gemakkelijker te leggen dan de onderliggende causaliteit. Dat er verbanden zijn, wil nog niet zeggen in welke richting het verband zich beweegt. Het was vooral verrassend om te zien dat de AI tot verbanden kwam, waar artsen niet aan gedacht hadden. Deze resultaten leidden tot meer medisch onderzoek op deze specifieke aspecten.*

Q : Jeroen, kun je vanuit jouw ervaring zeggen wat ongeveer de verhouding is van het whitebox en blackbox testen voor een AI product?

A : *Jeroen: Hier is geen eenduidig antwoord op te geven. Bij de ene organisatie zul je meer whitebox-testen zien dan bij de andere. Het dynamische, lerende gedrag is echter lastig te vangen in unit tests en zal dus meer bij de blackbox-tests uitkomen. Belangrijk is om altijd de testpiramide in het achterhoofd te houden en zo vroeg mogelijk in de cyclus te testen. Dus wanneer regels zich lenen om te unit testen, hier zeker op in zetten. Maar het voldoen aan verwachte resultaten over meerdere gebruikersgroepen heen, zal zich vooral in het blackbox-testen bevinden.*