Kwaliteit

Doorgaans heb je als A&O psycholoog te maken met “off the shelf” AI toepassingen. De ontwikkelaar/aanbieder heeft de documentatie die je nodig hebt om de kwaliteiten te beoordelen. Eventueel is nader onderzoek nodig.

Verzamel alle relevante gegevens om de kwaliteit van de AI te kunnen beoordelen. Hierbij denk je aan (beschrijvingen van) data (denk aan trainingsset en testset). De samenstelling en diversiteit in die datasets is belangrijk. De personen in de datasets moeten vergelijkbare diversiteit laten zien als de doelgroep waarop de AI gaat worden toegepast. Is gecontroleerd op bias en discriminatie? Zijn er bijeffecten waarmee rekening gehouden moet worden?

Verzamel gegevens over de werking van de algoritmen en over kwaliteitsfactoren van het resultaat, zoals consistentie (betrouwbaarheid), interpreteerbaarheid (begripsvaliditeit) en voorspellende kracht (criteriumvaliditeit).

Het is ook belangrijk dat de werking van de AI-toepassing transparant en uitlegbaar is. Dit geldt voor zowel de gebruikers als de doelgroep. Hoe zien gebruikers hoe de AI-toepassing werkt? Is de werking uit te leggen aan de doelgroep? De doelgroep, de client, dient goed, zorgvuldig en compleet geïnformeerd worden over gegevens die gebruikt zijn, het AI-resultaat, over hoe het tot stand is gekomen en wat dit resultaat betekent.

Bepaal of de kwaliteit van de AI te beoordelen is en zo ja, of deze voldoende is. Omdat de impact die een beslissing heeft op het leven van degene waarvoor de AI wordt toegepast groot kan zijn, zijn de eisen aan de kwaliteit van de gebruikte instrumenten navenant. Is dit hier het geval? Is de informatie compleet of mist er wat, waardoor er geen goed oordeel geveld kan worden?

Kwaliteit van de AI

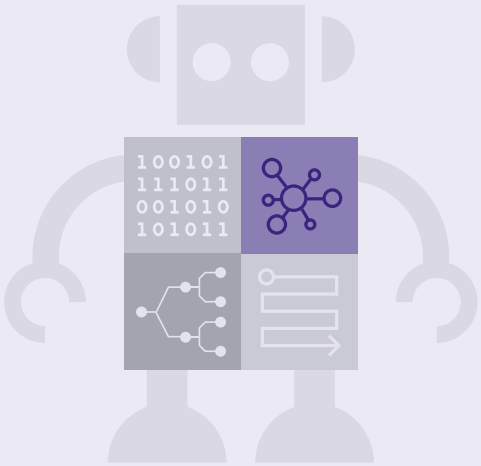
Verzamel alle relevante gegevens of doe er onderzoek naar:

- ☐ De testset die er is gebruikt
- ☐ De trainingsset die er is gebruikt
- ☐ Criteriumgegevens die zijn gebruikt
- ☐ Het gebruikte model (is het model - in combinatie met de data - solide, logisch, eerlijk?)
- ☐ De psychometrische kwaliteiten (hoe zit het met de theoretische basis, grootte van de steekproeven, betrouwbaarheid, normering, criterium en begripsvaliditeit, voorspellende waarde en bias?)
- ☐ Is er transparantie over de werking van de AI-toepassing?
- ☐ Is het aan de kandidaat uit te leggen welke gegevens zijn gebruikt en tot welke uitkomst dit leidt in de AI-toepassing?
- ☐ Zijn de (opgeleverde) gegevens compleet?

NVGzP NIP

Hoe beoordeel je de kwaliteit van een AI-toepassing?

We hebben eerder beschreven dat technische innovaties snel doordringen in de werving en selectie. De ontwikkeling van online assessment technieken, zoals gaming, virtual reality en AI, volgen elkaar snel op. Van elke techniek in de werving en selectie is de kwaliteit van belang.



Van een gedegen AI-toepassing mag je verwachten dat er verantwoordingsdocumentatie geleverd wordt waarin deze informatie staat. Soms ontstaat hier een spanningsveld: er zijn ontwikkelaars die een globale beschrijving van de werking als onderbouwing beschouwen en dit aanleveren als afdoende bewijs. Maar aan een AI-toepassing stel je vanzelfsprekend dezelfde kwaliteitseisen als aan andere technieken. Voor goede voorspellingen op basis van AI zijn heldere criteria, goede meetprocessen, controles op ongewenste effecten en goede data noodzakelijk (Tippins , 2021). Om de kwaliteit van de AI-toepassing te beoordelen, is documentatie van de leverancier nodig waaruit dit blijkt.

Specifiek voor het beoordelen van de kwaliteit van de werking van de AI-toepassing is het belangrijk om niet alleen informatie te hebben over de werking van de algoritmen, maar ook de gegevens waarmee de AI-toepassing 'getraind' is en de gegevens die de toepassing nu gebruikt. Bij de kwaliteitsbeoordeling moeten nadrukkelijk ook de gegevens over beide datasets worden meegenomen.

De Commissie Testaangelegenheden Nederland (COTAN) is de NIP-commissie die zich bezighoudt met de kwaliteit van tests en testgebruik in Nederland. Enerzijds doen zij dit door het beoordelen van de kwaliteit van instrumenten zoals tests, toetsen en vragenlijsten, anderzijds door het opstellen van standaarden over het gebruik van (psychologische) instrumenten, zoals de **Algemene Standaard Testgebruik (AST-NIP 2017)**. De richtlijnen in de AST-NIP zijn ook van toepassing op instrumenten die AI gebruiken. Daarnaast kunnen AI-toepassingen die (empirisch) onderzocht zijn, door de ontwikkelaar of leverancier ter beoordeling worden aangeboden aan de COTAN; dit raden wij ook aan. De COTAN geeft aan de hand van het **COTAN Beoordelingssysteem voor de kwaliteit van tests** een genuanceerd oordeel over de psychometrische kwaliteiten van het instrument, met daarbij in de Toelichting een bespreking van de mogelijkheden en beperkingen.

Is de AI-toepassing (nog) niet beoordeeld door de COTAN, dan zou je als psycholoog en hr-expert de criteria waar de COTAN tests op beoordeelt, zelf kunnen hanteren. Onder andere wordt een heldere meetpretentie, een bepaalde mate van betrouwbaarheid (consistentie en herhaalbaarheid), begripsvaliditeit (interpreteerbaarheid) en criteriumvaliditeit (voorspellende kracht) verwacht. Je kunt de ontwikkelaar of leverancier vragen de informatie zo aan te leveren dat je zelf kunt beoordelen in welke mate de toepassing aan deze eisen voldoet. Kanttekening is wel dat het beoordelingssysteem (nog) niet specifiek toegespitst is op AI-toepassingen.

Wat de hierboven door Tippins en collega's genoemde 'ongewenste effecten' betreft, zij doelden hier onder andere op discriminatie of bias. Het COTAN-beoordelingssysteem besteedt hier ook aandacht aan door middel van de *fairness* matrijs waarin beschreven staat welk onderzoek is uitgevoerd op dit gebied. Daarnaast is er de in opdracht van de Nederlandse overheid opgestelde **'Handreiking non-discriminatie by design'**, een leidraad voor het ontwikkelen en implementeren van AI-toepassingen die zo min mogelijk onbedoeld en ongerechtvaardigd onderscheid maken tussen groepen mensen. Hierin worden suggesties gegeven om discriminatie te vermijden bij de probleemdefinitie, dataverzameling, datavoorbereiding, modellering, implementatie en evaluatie van de AI-toepassing. De handreiking is opgesteld voor ontwikkelaars, en kan ook ter beoordeling van de verantwoordingsdocumentatie van de AI-toepassing gebruikt worden.

Ook internationaal wort er gewerkt aan kwaliteitsbeoordelingssystemen voor AI. De **ATP (Association of Test publishers)** en **ITC (International Test Commission)** werken momenteel aan specifieke richtlijnen voor assessments met een technologische basis en/of in een digitale omgeving, deze worden eind 2022 verwacht. Maar ook deze zullen, net als die van de COTAN, niet specifiek op AI zijn toegespitst.

De APA (American Psychological Association) heeft een bruikbaar model opgesteld om AI-systemen te auditeren op fairness en bias (**Landers & Behrend, 2022**). Het bevat vragen die kunnen worden gesteld over de onderzoeksopzet en het algoritme (de data, het onderzoeksmodel, voorspellingsresultaten), de effecten op de omgeving (voor de geselecteerden, niet-geselecteerden, de opdrachtgever, de maatschappij) en de mate waarin voldaan wordt aan de wettelijke eisen, standaarden en ethiek.

Tabel *Components of AI System to Be Audited*

Component	Questions to ask	Applied to focal example
Components relating to models		
Input data	How were input data collected in terms of population and research design? How did these factors affect data quality?	Were the input data collected from job incumbents? How are incumbents different from applicants? Is there range restriction on variables of interest?
Model design	What drove initial model choices (e.g., criterion, predictor set, algorithm)? Were they informed by theory or empirically derived? If empirically derived, from what data (and of what quality were those data)?	Were the input data collected from job incumbents? How are incumbents different from applicants? Is there range restriction on variables of interest?
Model development	Once the initial model was created, how was it refined? What approaches were taken, and what likely effect did these approaches have?	Is every decision about every model created during the development process fully documented? When were models discarded and why?
Model features	How were the raw input data engineered into model features? Was this process conceptually or empirically driven? What alternative feature engineering approaches were explored?	What specific natural language processing technique were used? What evidence is there of quality speech-to-text conversion? What facial features emerged from analyzing video? What biases were explored from these engineering processes?
Model processes	How does the model use inputs to generate scores? How were alternative approaches explored and evaluated?	What stress tests were conducted, and what types of bias were investigated? Did these tests result in changes to the model, and if so, how and why?
Model outputs	How was the quality of predictions generated by the model evaluated, such as for psychometric reliability and validity? How was cross-validation conducted, and was it appropriate given claims about model generalizability?	Are scores consistent over time and upon multiple administrations (i.e., reliability evidence)? Is there evidence that they reflect the predicted constructs they claim to (i.e., validity evidence)? Do they show differences among classes of interest (e.g., race, gender, color, national origin, religion, disability, age) and combinations of classes?
Components relating to information and perceptions		
First-party interpretation	Does all messaging from the algorithm developer logically, honestly, and transparently follow from answers developed elsewhere in this audit?	Does the developer claim the model predicts job performance? What evidence in the audit forms the basis of this claim? Is anything exaggerated? Are important details left out?
Second-party effects	Who is directly affected by the use of the algorithm, and how have their outcomes and reactions been assessed? What is the relative impact of acting upon false positives versus false negatives on second parties?	How do nonselected applicants react to the news that the algorithm did not assign them a high enough score to be selected? What information is communicated to them, and how do they evaluate that information?
Third-party understanding	How have perceptions and evaluation by outside observers been assessed and incorporated? Have outside regulatory groups and community organizations been consulted?	How do experts in employment law view the documentation and performance of the algorithm? How does the public view this use of algorithms?
Meta-components		
Cultural context	Has the broader cultural context in which the algorithm will be used been considered? Have members of the community participated in the design of systems that will affect them?	Do power differentials exist between designers, employers, and job candidates? Have cultural assumptions been made? Will development decisions in one culture be applied to another, and if so, how has the development process been adjusted to prevent cross-cultural application challenges?
Respect	Is the algorithm being used in a way that conforms to generally accepted ethical standards, such as those in the Standards, the SIOP Principles, the OECD Principles and the UGAI?	What ethical standards do the developers claim to have followed in development? Is there evidence of decisions made following that ethical framework?What evidence is there that individual fairness has been a priority in development?
Research designs	How do the research designs (including sampling, experimental design, variable choices, analysis, and interpretation) of any studies conducted to support a claim affect the validity of conclusions?	For every claim that appears to be based upon empirical observation, does the study design support the claims made? Were all design decisions defensible from the perspective of modern methodological research? What impact might they have had on the validity of drawn conclusions?

AI= artificial intelligence; SIOP = Society for Industrial and Organizational Psychology; OECD = Organisation for Economic Co-operation and Development; UGAI = Universal Guidelines for Artificial Intelligence.