AI blindspots

.AGORIA

Kenniscentrum
Data & Maatschappij

Deze AI Blindspots-kaartenset is gebaseerd op AI Blindspot, dat beschikbaar is onder een Creative Commons Attribution 4.0 internationale licentie.

Het Kenniscentrum Data & Maatschappij paste de originele kaartenset aan de Vlaamse context aan, om de ontwikkeling van betrouwbare AI in Vlaanderen te ondersteunen. Agoria maakte de aanpassingen om deze aan te bieden voor het belgische eco-systeem.

AGORIA

Kenniscentrum Data & Maatschappij

Wat zijn AI Blindspots, en hoe detecteer je ze?

AI Blindspots zijn dingen die vaak over het hoofd worden gezien vóór, tijdens of na de ontwikkeling van een AI-systeem. Ze komen voort uit vooroordelen, vooringenomenheid en structurele ongelijkheden in de maatschappij.

De ongunstige gevolgen van AI Blindspots voorspellen is een hele uitdaging. Je kunt die gevolgen wel afzwakken door ze proactief te detecteren en passend te handelen.

Deze kaartenset kan je helpen mogelijke AI Blindspots te ontdekken door op voorhand na te denken over de gevolgen van beslissingen en acties.

Elke kaart bevat een aantal vragen over mogelijke blindspots, geeft je een voorbeeld uit de praktijk dat het belang van deze blindspot verduidelijkt en reikt je enkele tools en tips aan om blindspots te detecteren en te beperken.

De kaartenset bevat ook een joker zodat je andere AI blindspots kan toevoegen die jij en je team ontdekken.

Deze kaartenset is gebaseerd op AI Blindspot van Ania Calderon, Dan Taber, Hong Qu en Jeff Wen, die hun kaartenset ontwikkelden tijdens het 2019 Assembly program van het Berkman Klein Center en het MIT Media Lab.

Het Kenniscentrum Data & Maatschappij paste de originele kaartenset aan de Vlaamse context aan, om de ontwikkeling van betrouwbare AI in Vlaanderen te ondersteunen. Agoria maakte de aanpassingen om deze aan te bieden voor het belgische eco-systeem.

De originele kaartenset identificeert drie fases (planning, ontwikkeling en implementatie). Deze kaartenset focust louter op de eerste fase, nl. de planning. In de toekomst wordt deze kaartenset verder uitgewerkt met kaarten voor de andere twee fases.

DOEL

Bij de start van een AI-project bepaal je het doel van je AI-systeem. Daarbij praat je met stakeholders, experts en je team om je doel en het probleem waarop je AI-systeem een antwoord moet bieden, duidelijk af te lijnen.

FASE: PLANNING

HEB JE HIERAAN GEDACHT?

A. Heb je het probleem en het resultaat waarvoor je optimaliseert **duidelijk uitgedrukt**?
B. Is de **tool geschikt** om dit resultaat te bereiken?
C. **Erkennen** alle geraadpleegde en betrokken **stakeholders** dit als een belangrijk probleem?
D. Heb je nagedacht over de **voordelen en nadelen** van je AI-systeem voor elke stakeholder?
E. Hoe zal je **garanderen** dat het AI-systeem **gericht blijft op het gestelde doel**?

ZO MOET HET NIET

Een bedrijf introduceerde een AI-systeem om de productie te versnellen. Indirect had dit als gevolg dat werknemers hun bonussen kwijtraakten. Dit had vermeden kunnen worden door met de vakbonden, als betrokken stakeholder in het project, te praten over manieren om de snelheid te verhogen zonder dat de bonus verloren gaat.

TOOLS & TIPS

A&B: template voor probleemstelling
B: cursus over 'machine learning' (Google)
C: andere applicaties van je 'machine learning' toepassen op je case: klopt het nog?
D: stakeholders identificeren en valideren

DATA-EVENWICHT

Data-evenwicht houdt in dat je bent nagegaan of je data representatief is en dat je hebt nagedacht over hoe je onevenwicht in je data kan reduceren.

FASE: PLANNING

.AGORIA Kenniscentrum Data & Maatschappij

HEB JE HIERAAN GEDACHT?

A. Wat is de **minimaal haalbare gegevensverzameling** die je volgens domeinexperts nodig hebt?
B. Wat of wie kan **uitgesloten** worden **in je data**?
C. Welke gevolgen hebben **beperkingen van je data** op de representativiteit van je model en de acties die door je model worden ondersteund?
D. Als je **data niet in evenwicht** is, hoe kan je deze beperking dan minimaliseren?
E. Kan je, rekening houdend met je data, de case of persoon beschrijven waarvoor je **voorspellingen het meest onbetrouwbaar** zullen zijn?

ZO MOET HET NIET

Na de release van de enorm populaire Pokémon Go merkten verschillende gebruikers op dat er minder Pokémon-locaties waren in voornamelijk zwarte wijken. Dit kwam omdat de makers van de algoritmen er niet in slaagden om een gevarieerde trainingsset te gebruiken, en geen tijd doorbrachten in deze buurten.

TOOLS & TIPS

- A: interview met een domeinexpert
- B, C & D: Data Collection Bias Assessment, Aequitas
- E: maak een persona van de onzichtbare vrouw/man

DATABESTUUR & PRIVACY

Vragen over databestuur en de impact op de privacy van de betrokkenen van wie de persoonsgegevens door het AI-systeem worden verwerkt, horen bij de voorbereiding van je AI-project. Door het niveau van de toegang tot data te bepalen en de informatiestroom te beschrijven, kun je de rechten van je betrokkenen beter beschermen.

FASE: PLANNING

.AGORIA Kenniscentrum Data & Maatschappij

HEB JE HIERAAN GEDACHT?

A. Kan je de **data rechtmatig verwerken of hergebruiken**?
   - Als je de data hergebruikt, is het doel hetzelfde?
   - Zijn er passende contractuele afspraken?
   - Kan je de data verwerken of hergebruiken op basis van toestemming of andere gronden?

B. Verzamel je **gevoelige gegevens**?

C. Heb je **speciale procedures om je data te beveiligen**?

D. Wie krijgt **toegang tot de (verzamelde) data** (intern en extern)?

E. Kan je **voldoen aan de rechten van de betrokkenen** op grond van de AVG (GDPR)?

ZO MOET HET NIET

Een Brits ziekenhuis dat samen met Deepmind werkt aan een AI-toepassing voor het opsporen en diagnosticeren van nierschade werd beboet voor het overtreden van de regels voor persoonsgegevens. Het had persoonlijke gegevens van 1,6 miljoen patiënten doorgegeven zonder dat zij daarover voldoende waren geïnformeerd.

TOOLS & TIPS

- Data Protection Impact Assessment
- Gegevensstromen in kaart brengen
- Contact met gegevensbeschermingsdeskundigen

SAMENSTELLING VAN HET TEAM

Het is moeilijk om inzicht te krijgen in mogelijke (ethische) kwesties als je niet op de hoogte bent van vooroordelen binnen je team. Dergelijke blindspots kun je alleen vermijden door ze bloot te leggen.

FASE: PLANNING

SAMENSTELLING VAN HET TEAM

HEB JE HIERAAN GEDACHT?

A. Heb je gedacht aan **vooroordelen** binnen je team?
B. Is je **team divers en multidisciplinair**, of is het op de hoogte van het probleemgebied waarvoor je een oplossing probeert te vinden?
C. **Wie moet je uitnodigen** om dit verkeerde idee bloot te leggen?

ZO MOET HET NIET

De software van Google die foto's automatisch in categorieën onderbrengt, verwarde zwarte mensen soms met gorilla's. Als de dienst getest was door leden van het team met een donkere huid, had deze uitkomst drastisch beperkt kunnen worden.

TOOLS & TIPS

- A: impliciete associatie test
- B: bezoek ter plaatse, 'empathy map', persona, ...

GRENSOVER-SCHRIJDENDE EXPERTISE

Misschien ben je een expert in 'machine learning', maar niet in het domein waarop je 'machine learning' toepast. Dit is geen probleem als je kunt rekenen op een expert die je kan vertellen welke typische uitschieters, belangrijke variabelen of gebruikelijke praktijken van invloed kunnen zijn op je data.

FASE: PLANNING

HEB JE HIERAAN GEDACHT?

A. Heb je met **domeinexperts** besproken wat de **minimaal haalbare gegevensverzameling** is die je nodig hebt zodat je AI-systeem zijn doel kan bereiken?
B. Welke **variabelen** zijn van **essentieel belang** voor je probleem?
C. Heb je een **expert** ingeschakeld om je te helpen de **resultaten van je algoritme te beoordelen**?
D. Heb je een **expert** ingeschakeld om te begrijpen wat de **impact van je algoritme zou moeten zijn**?

ZO MOET HET NIET

Een nieuw algoritme moest bepalen wie op een spoedafdeling meteen zou worden gecontroleerd op een longontsteking. Volgens het algoritme hadden mensen met astma geen onmiddellijke zorg nodig. Experts gingen hier niet mee akkoord omdat astmapatiënten dringend worden behandeld op een spoedafdeling. De experts zagen dat dit was gebaseerd op verkeerde veronderstellingen door het AI-systeem. Uit de trainingsdata, bleek dat astmapatiënten het minst lang op de spoedafdeling bleven. Daarom besloot het AI-systeem dat zij voor de doeltreffendheid van de spoedafdeling niet belangrijk waren.

TOOLS & TIPS

- Interview of focusgroep met expert(en)
- Workshop over technische en systeemvereisten

KANS OP MISBRUIK

Je AI-systeem is bedoeld om de wereld beter te maken. Als je echter alleen focust op de positieve aspecten, verlies je misschien uit het oog hoe het systeem ook schade kan aanrichten. Voorkomen is altijd beter dan genezen. Bedenk dus wat een echt kwaadwillige partij kan doen met je applicatie.

FASE: PLANNING

KANS OP MISBRUIK

FASE: PLANNING

HEB JE HIERAAN GEDACHT?

A. Hoe kan het AI-systeem **onethisch** worden gebruikt?
B. Wat zouden de **gevolgen** zijn van dergelijk onethisch gebruik van je AI-systeem?
C. Wie heb je ingeschakeld om de **onderliggende sociale motivaties en dreigingsmodellen** te begrijpen?
D. Wat is je **afzwakkingsstrategie** voor het geval je AI-systeem onethisch wordt gebruikt?
E. Wat doe je als je algoritme **onethisch gedrag** ontwikkelt?
F. Wat zijn de **belangrijkste ethische principes** waarmee je AI-systeem moet rekening houden?

ZO MOET HET NIET

In 2016 bracht Microsoft een Twitter-chatbot uit met de naam Tay. Binnen 24 uur klonk Tay helemaal anders, omdat ze op basis van de tweets die naar haar werden verstuurd had geleerd een racistische twittergebruiker te worden. Microsoft trok de chatbot daarom al snel terug.

TOOLS & TIPS

- Maak scenario's om de schadelijke en onethische praktijken van je systeem te begrijpen, en bekijk de gevolgen ervan voor onschuldige niet-betrokken persona's.
- Vraag raad aan experten uit sociale en rechtswetenschappen

JOKER KAART

FASE: PLANNING

JOKER KAART **FASE: PLANNING**

HEB JE HIERAAN GEDACHT?

ZO MOET HET NIET

TOOLS & TIPS