Bijlage 3. Considerati onderzoek: onvoorziene effecten in algoritmes

1. Samenvatting

Het effectenonderzoek van Considerati[1] is tot stand gekomen in reactie op de motie van de Kamerleden Van den Berg en Wörsdörfer waarbij de regering wordt verzocht te onderzoeken in wat voor situaties ongewenste effecten bij algoritmes zich voor kunnen doen en hoe die voorkomen kunnen worden.[2]

De hoofdvraag van het onderzoek luidt: Wat zijn mogelijke onvoorziene effecten van de inzet van zelflerende algoritmes door bedrijven en (voor en door) consumenten waarvan niet duidelijk is hoe zij tot een besluit komen en hoe kunnen deze effecten geïdentificeerd, gewogen en indien ongewenst gemitigeerd worden?

Deze vraag wordt beantwoord aan de hand van literatuuronderzoek en interviews en vier casestudies:

- **het bepalen van prijzen met behulp van algoritmes (algorithmic pricing**). Het dynamisch prijzen van producten of diensten op basis van de inzichten van een algoritme kan grote economische meerwaarde hebben, bijvoorbeeld omdat minder wordt verspild en vraag en aanbod elkaar optimaal weten te vinden. Tegelijkertijd kan dynamisch prijzen mogelijk ook leiden tot oneerlijke prijsdiscriminatie of tot collusie tussen marktpartijen. Zonder inzicht in de totstandkoming van de prijzen en de factoren die daarbij een rol spelen zijn deze ongewenste effecten niet te detecteren.
- **het beoordelen van de betrouwbaarheid van een persoon (fraude, kredietwaardigheid).** Een concrete case study voor de toepassing van algoritmen voor risico- inschatting is het toetsen van kredietwaardigheid van personen. De bank analyseert bijvoorbeeld op basis van allerlei gegevens over de kredietaanvrager of hij/zij in de toekomst een lening kan terug betalen.
- **het beoordelen van de geschiktheid van een persoon (HR analytics).** Deze casus betreft het gebruik van algoritmes voor het monitoren of beoordelen van sollicitanten en medewerkers. Het gebruik van algoritmische besluitvorming kan tot betere HR-beslissingen leiden, maar kan ook de verhouding tussen werkgever en werknemer ingrijpend veranderen en discriminatie in de hand werken.
- **gepersonaliseerde advisering (fysieke & geestelijke gezondheid).** Dit heeft betrekking op het gebruik van algoritmes (in applicaties als nutrition trackers) om consumenten te helpen (betere) beslissingen te nemen over hun fysieke en/of mentale gezondheid. Deze applicaties registreren de gegevens over het subject, geven advies of verwijzen door naar een professionele hulpverlener of arts. Deze toepassingen ondersteunen mensen om betere keuzes te maken in hun leven en kunnen de zorg ontlasten. Ze roepen echter ook vragen op over de autonomie van de consument, en de aansprakelijkheid voor foutieve adviezen.

Belangrijkste conclusies
Het onderzoek concludeert dat de toepassing van (zelflerende) algoritmes in alle onderzochte cases kan bijdragen aan de efficiëntie en accuraatheid van besluitvorming. Onvoorziene effecten ontstaan vaak niet door de algoritmes zelf, maar door een mismatch tussen het algoritme en de context waarbinnen het wordt toegepast. Een verkeerde toepassing kan leiden tot onvoorziene effecten die door hun onvoorspelbaarheid doorgaans negatief van aard zijn. De onvoorziene effecten kunnen echter door een goed doordachte en zorgvuldige toepassing van kunstmatige intelligentie beperkt worden.

<u>Grondoorzaken van onvoorziene effecten</u>
Het onderzoek onderscheidt drie 'grondoorzaken' voor het optreden van onvoorziene effecten in de context van de toepassing van (zelf)lerende algoritmes:

1. Er is een onvolledig of verkeerd begrip van de probleemruimte.

[1] Considerati onderzoek: onvoorziene effecten in algoritmes: https://www.rijksoverheid.nl/documenten/rapporten/2020/09/14/onvoorziene-effecten-van-zelflerende-algoritmen
[2] Motie Van den Berg/Wörsdörfer over een onderzoek naar de ongewenste effecten van algoritmes 21501-33, nr. 748 https://www.tweedekamer.nl/kamerstukken/detail?id=2019Z03423&did=2019D07342

2. Het zelflerende algoritme is niet goed toegerust om om te gaan met de complexe omgeving waarbinnen het wordt ingezet.
3. Het zelflerende algoritme wordt niet goed ingepast in een bredere (socio-technische) context

Probleemruimte
De probleemruimte is de omgeving waarbinnen kunstmatige intelligentie wordt ingezet en waarbinnen het specifieke doelstellingen moet bereiken. Een probleem wordt door het algoritme opgelost op basis van de gegevens die het krijgt aangeleverd. Deze gegevens, en hun onderlinge relaties, zijn een omschrijving van de werkelijkheid. Door af te bakenen welke gegevens het algoritme kan gebruiken worden de keuzemogelijkheden die het heeft bepaald.

Context afhankelijk
De mate van onvoorziene gevolgen van algoritmes en de impact daarvan, is in belangrijke mate afhankelijk van de vraag hoe ontwikkelaars en gebruikers omgaan met de bovenstaande grondoorzaken. Die spelen nu al een rol, maar huidige toepassingen van zelflerende algoritmes zijn nog weinig complex en hun onvoorziene effecten zijn volgens de onderzoekers daarom nu nog beperkt. Het risico op onvoorziene effecten op de langere termijn zal naar verwachting veel groter zijn als de ontwikkeling van zelflerende algoritmen doorzet.

Emergent gedrag
Op de langere termijn kan het vraagstuk van emergent[3] gedrag van algoritmes als grondoorzaak relevant worden. Emergent gedrag ontstaat als er interactie is tussen zelflerende algoritmes en deze zelf nieuwe strategieën ontwikkelen op basis van veranderde omstandigheden die niet door de ontwikkelaar of gebruiker waren voorzien. Op dit moment lijken er nog weinig situaties te zijn waar emergent gedrag naar voren komt.

Maatregelen
Om de kans op onvoorziene effecten van (zelflerende) algoritmes te verkleinen noemt het onderzoek een aantal maatregelen om de grondoorzaken van die onvoorziene effecten weg te nemen. Het betreft veelal organisatorische en technische maatregelen die bedrijven kunnen nemen om ervoor te zorgen dat de algoritmes doordacht en zorgvuldig worden ontwikkeld en toegepast. Ook worden maatregelen voorgesteld met betrekking tot toezicht en consumentenbescherming.

Organisatorische maatregelen
Veel van de zorgen omtrent algoritmische besluitvorming kunnen worden weggenomen door het goed definiëren en begrijpen van het probleem dat het model moet oplossen en de data die nodig zijn om dit probleem goed te modelleren. Hiervoor zijn verschillende modellen bruikbaar zoals het Cross-Industry Standard Process for Data Mining (CRISP-DM). Ook moet de toepassing van het model in de praktijk nauwkeurig gemonitord worden, met passende maatregelen op het gebied van governance en risk management.

Daarnaast is het belangrijk dat gebruikers van algoritmes een besef hebben van de risico's van de toepassing van algoritmische modellen. Naast bewustwording binnen een organisatie en bij ontwikkelaars is daarvoor ook een proces nodig voor het inschatten van de risico' en hoe hiermee om te gaan. Daarin kunnen impact assessments waarbij ook wordt gekeken naar de context waarin het algoritme wordt toegepast een belangrijke rol spelen.

Technische maatregelen
Technische maatregelen zijn met name gericht op het juist samenstellen van (trainings)data, het testen en valideren van modellen en het inzichtelijk of transparant maken van de al dan niet oneerlijke uitkomsten van algoritmes. Daarvoor zijn bestaande methoden en technieken beschikbaar, die in combinatie met organisatorische maatregelen moeten worden toegepast.

(Extern) toezicht
Intern en extern toezicht op de ontwikkeling en toepassing van zelflerende algoritmes is vooral van belang bij toepassingen waar een potentieel grote impact op de rechten en vrijheden van personen, organisaties of groepen valt te verwachten. Transparantie en uitlegbaarheid zijn cruciale voorwaarden daarvoor. Om goed toezicht te kunnen houden op zelflerende algoritmen en onvoorziene effecten zoveel mogelijk te voorkomen moeten bedrijven de ontwikkeling en het gebruik van algoritmes kunnen verantwoorden.

Het beschermen van de rechtspositie van subjecten
Omdat het subject van een algoritmisch model (een werknemer, een consument of een bedrijf) doorgaans geen invloed kan uitoefenen op de werking van een model, zijn ook maatregelen ter bescherming of versterking van de rechtspositie van het subject relevant. Het gegevensbeschermingsrecht, het aansprakelijkheidsrecht, het consumentenrecht en meer indirect het mededingingsrecht zijn daarbij aan de orde. Daarnaast is bewustwording over algoritmische besluitvorming en de effecten daarvan voor subjecten noodzakelijk. Dit kan onder andere door transparantie over algoritme-toepassingen te vergroten.

2. Appreciatie

Het Considerati onderzoek geeft een goede en begrijpelijke omschrijving van (de verschillende vormen van) artificiële intelligentie (AI). Het geeft een duidelijke uitleg over (zelflerende) algoritmes en hoe onvoorziene effecten tot stand komen.

De vier case studies geven een duidelijk beeld van de wijze waarop algoritmes nu al worden ingezet bij bedrijfsprocessen en van hun onvoorziene effecten. Een belangrijke bevinding is dat het niet alleen gaat om de techniek van het algoritmes, maar de wijze waarop en de context waarin ze worden toegepast, die de bron vormt van ongewenste en onvoorziene effecten.

Om de onvoorziene effecten van algoritmes te beperken stellen de onderzoekers verschillende technische en organisatorische maatregelen voor; veelal reeds bestaande methodologieën en richtlijnen. Het kabinet erkent de waarde daarvan en ondersteunt initiatieven die daar aan bijdragen. Op Europese niveau zijn Europese richtsnoeren voor betrouwbare AI ontwikkeld en binnen de Nederlandse AI Coalitie is een werkgroep actief op het gebied van human centric AI. De Nederlandse AI Coalitie draagt ook bij aan het ontwikkelen en verspreiden van kennis over verantwoord gebruik van AI.

Het onderzoek concludeert dat de risico's op onvoorziene effecten contextafhankelijk zijn en met bestaande technische en organisatorische maatregelen beperkt kunnen worden, op de kortere termijn althans. Dit sluit aan bij de door Nederland gesteunde plannen in het AI Witboek[4] van de Europese Commissie voor een 'lerende aanpak', waarbij wordt gekeken bij welke risicovolle sectoren en risicovolle toepassingen de huidige (EU-)wetgeving niet of onvoldoende bescherming biedt.

Binnen deze lerende aanpak wordt ingezet op de ontwikkeling van AI door onderzoek, experimenten en pilots binnen de kaders van de bestaande wet- en regelgeving. Ze ontstaat vroegtijdig zicht op mogelijke problemen met de ontwikkeling en acceptatie van AI en kunnen waar nodig aanvullende maatregelen genomen worden. De voorgestelde technische, organisatorische en toezicht maatregelen bieden daartoe veel mogelijkheden.

Het onderzoek benadrukt dat veel vormen van regulering kunnen ondersteunen bij het mitigeren van de onvoorziene effecten via het gegevensbeschermingsrecht, het aansprakelijkheidsrecht, het consumentenrecht en meer indirect het mededingingsrecht. In aanvulling op bestaand beleid en regulering komt de Europese Commissie op basis van het Witboek AI begin volgend jaar met een pakket aan beleidsvoorstellen rondom AI, waarin naar alle waarschijnlijkheid aansprakelijkheid, veiligheid en eventuele risico's per toepassing aan de orde komen. Het is op dit moment onduidelijk of het een regulering of richtlijn wordt. Het kabinet blijft sturen op een lerende aanpak, gericht op risicovolle sectoren en toepassingen.

Het kabinet onderschrijft het belang van goede consumentenbescherming. Binnen de Europese richtsnoeren voor betrouwbare AI is al aandacht voor transparantie en het kabinet steunt de in het AI witboek aangekondigde plannen van de Europese Commissie om concrete eisen aan de transparantie van AI te stellen. Het kabinet gaat bovendien nader verkennen welke maatregelen het beste kunnen bijdragen aan het tegengaan van eventuele negatieve effecten van algoritmen op consumenten.

[4] Kabinetsappreciatie witboek over Kunstmatige intelligentie, april 2020. Zie: https://www.rijksoverheid.nl/documenten/kamerstukken/2020/04/20/aanbieding-kabinetsappreciatie-witboek-kunstmatige-intelligentie