Rathenau Instituut

Zeven aandachtspunten voor de AI-verordening

Bericht aan het parlement

Artificiële intelligentie (AI) wordt steeds breder toegepast in de maatschappij: van gezichtsherkenning op smartphones, tot het stellen van vroegtijdige medische diagnoses. AI kan ook mensenrechten schaden, zoals de rechten op privacy en non-discriminatie. De Europese Commissie heeft daarom een wetsvoorstel gepresenteerd: de AI-verordening. Een belangrijke stap, maar om te zorgen dat het voorstel erin slaagt om te komen tot AI-systemen die in lijn zijn met mensenrechten, is meer nodig. In dit Bericht aan het parlement benoemt het Rathenau Instituut zeven aandachtspunten voor de behandeling van de verordening in de Tweede Kamer. De belangrijkste boodschap is dat aanscherping van de voorgestelde verordening nodig is om ernstige risico's daadwerkelijk te beteugelen. Bovendien is het zaak om niet slechts de risico's kritisch te wegen, maar evengoed de maatschappelijke opbrengst van AI-systemen te beoordelen.

De AI-verordening

De Europese Commissie en het kabinet zien AI als een sleuteltechnologie voor de toekomst: essentieel om maatschappelijke opgaven aan te pakken – van toekomstbestendige zorg tot duurzame energie – en essentieel voor de Nederlandse en Europese economie. Tegelijkertijd erkennen zij dat AI-systemen maatschappelijke uitdagingen opleveren. De systemen kunnen leiden tot onnavolgbare besluiten en discriminerende uitkomsten. In Nederland kwam deze discussie op scherp te staan door de toeslagenaffaire. In de samenleving zijn er daarnaast zorgen over biometrische systemen, die gebruikmaken van unieke lichaams- of gedragskenmerken en mensen in de openbare ruimte op afstand kunnen identificeren. Hierdoor wordt de anonimiteit van burgers goeddeels opgeheven. Ook de opkomst van deepfakes en deepnudes baart zorgen, net als de groeiende mogelijkheden om mensen te profileren en te beïnvloeden.

Het kabinet en de Europese Commissie willen daarom investeren in ethische AI: AI-systemen die publieke waarden en mensenrechten, zoals privacy en gelijke behandeling, in acht nemen. De Europese Commissie presenteert de AI-verordening. Het wetsvoorstel stelt geharmoniseerde regels op voor ‘het in de handel brengen, in gebruik stellen en gebruiken’ van AI-systemen.

De Commissie oordeelt dat een aantal AI-systemen mensenrechten te zwaar inperken en stelt vier typen verboden voor. Verder stelt het wetsvoorstel aan AI-systemen met een hoog risico op mensenrechtenschending specifieke eisen. En voor bepaalde AI-systemen, zoals een chatbot of deepfakes, komt een transparantieverplichting. Het moet duidelijk zijn dat je met een chatbot praat, of dat je iets ziet dat nooit echt gebeurd is. Op het niet naleven van deze voorschriften staan boetes. De boete voor het niet naleven van een verbod is bijvoorbeeld gesteld op 30.000.000 euro of tot 6% van de jaarlijkse wereldwijde omzet van een onderneming.

In dit Bericht aan het parlement noemen we zeven aandachtspunten voor de behandeling van de AI-verordening in de Tweede Kamer:

1. Beperk de mogelijkheden voor onbewuste beïnvloeding;
2. Bescherm kwetsbare groepen beter tegen onbewuste beïnvloeding;
3. Scherp het verbod op sociale scores door overheden aan;
4. Verbied biometrische identificatie in de openbare ruimte – niet alleen voor opsporing;
5. Vul de lijst van hoog-risico-systemen aan;
6. Beoordeel de maatschappelijke opbrengst van AI-systemen;
7. Transparantie is belangrijk, maar onvoldoende.

1. Beperk de mogelijkheden voor onbewuste beïnvloeding

Het eerste verbod (artikel 5a) dat de AI-verordening voorstelt betreft ‘subliminale’ beïnvloedingstechnieken die leiden tot fysieke of psychologische schade. Via AI wordt op het internet continu zeer gevoelige informatie over mensen verzameld, bijvoorbeeld

over iemands seksuele of politieke voorkeuren. Deze informatie kan worden gebruikt om mensen te misleiden. Het verbod van artikel 5a poogt daarom de menselijke autonomie te beschermen. Maar het is niet helder wat er onder subliminale technieken wordt verstaan. Welke vormen van beïnvloeding perkt het voorgestelde verbod precies in?

De Europese Commissie houdt verder bepaalde vormen van onbewuste beïnvloeding buiten de AI-verordening, zoals de beïnvloeding van consumenten en politieke microtargeting, terwijl de Europese Commissie onderkent dat ook zo de menselijke autonomie wordt ingeperkt. Voorlopig blijven er dus vele mogelijkheden bestaan voor onbewuste online beïnvloeding. Die zullen lang niet altijd resulteren in fysieke of psychologische schade. Maar er zijn ook andere vormen van schade, bijvoorbeeld gezondheidsschade of maatschappelijke schade door beïnvloeding van het publieke debat. Het Rathenau Instituut roept daarom op tot meer transparantie over de gebruikte methoden voor targeting en de gebruikte data.

Bovendien is de verwachting dat met de opkomst van immersieve technologie (Augmented Reality, Virtual Reality en spraaktechnologie) de beïnvloeding zich uitbreidt naar de fysieke wereld. Op pleinen, straat, winkels en andere openbare ruimtes zullen AR-apps op slimme schermen, smartphones of brillen, ruimte bieden om extra informatie of reclame te tonen. Daarmee groeien de mogelijkheden voor onbewuste sturing van de beelden en geluiden die we waarnemen, en welke niet. Welke vormen van onbewuste beïnvloeding vinden we in daarin wel en niet toelaatbaar? Heldere afspraken zijn nodig. De DSA en DMA, eveneens wetsvoorstellen in voorbereiding in Europa, bieden hiertoe mogelijkheden. Ook daar roept het Rathenau Instituut op tot aanscherpingen van de huidige voorstellen.

Bepaal welke vormen van onbewuste sturing, online en offline, ongewenst zijn, en stel daar nadere eisen aan, bijvoorbeeld ten aanzien van politieke microtageting.

2. Verbreed de bescherming van kwetsbare groepen tegen onbewuste beïnvloeding

Met het tweede verbod (artikel5b) wil de Commissie kinderen en mensen met een beperking beschermen tegen schadelijke beïnvloeding. Het is opmerkelijk dat het wetsvoorstel alleen deze twee kwetsbare groepen benoemt. Het recht op een gelijke behandeling benoemt meer groepen op basis van geslacht, etnische achtergrond, religie, politieke opvattingen, leeftijd of seksuele voorkeur. Al deze groepen kunnen met AI-technieken worden beïnvloed, wat kan resulteren in fysieke of psychologische schade. Zo laat onderzoek bijvoorbeeld zien dat vrouwen onevenredig hard worden getroffen door seksueel georiënteerde deepfakes. Naar schatting betreft 90-95% van alle deepfakes ‘nonconsensual’ pornografie, waarvan ca 90% specifiek gericht is op vrouwen.

Overweeg het verbod op de bescherming van kwetsbare groepen tegen onbewuste beïnvloeding uit te breiden en gelijk te trekken met bestaande wetgeving voor gelijke behandeling.

3. Scherp het verbod op sociale scores door overheidsinstanties aan

Het derde verbod (artikel5c) moet voorkomen dat mensen, of groepen mensen, nadelig worden behandeld door overheden, doordat ze op basis van hun gedrag of kenmerken zouden worden geclassificeerd met ‘een sociale score’. De score mag niet leiden tot 1) een nadelige behandeling van personen in sociale contexten die ongerelateerd zijn aan de context waarin de data oorspronkelijk verzameld is, of tot 2) een nadelige behandeling van personen die ongerechtvaardigd of onevenredig is met hun sociale gedrag of de ernst hiervan. De Europese Commissie wil hiermee voorkomen dat een sociaal scoresysteem, zoals China dat kent, in Europa kan worden toegepast. Maar de Europese Commissie kijkt ook dichter bij huis. Het door de Nederlandse rechter verboden risicoprofileringsysteem SyRi wordt door de Europese Commissie bijvoorbeeld expliciet genoemd als ongewenst voorbeeld van sociale scoring.

Het doel van de Commissie is niet om sociale scores onmogelijk maken, maar om excessen te voorkomen. De vraag is of het verbod daar in slaagt, gezien de abstracte formuleringen. Want wanneer zijn sociale contexten precies ‘ongerelateerd’ aan de oorspronkelijke verzamelcontext? Kan er met een wetsvoorstel een legitieme relevante context worden gecreëerd? En wordt de sociale score ingezet om mensen op maat te helpen, of brengt deze score hen verder in de knel? In Nederland zijn deze vragen actueel, gezien de toeslagenaffaire en nu de opvolger van SyRI in de Eerste Kamer voorligt.

De Tweede Kamer kan de minister vragen om te verhelderen hoe excessen worden voorkomen, zowel met het oog op de mogelijke nadelige behandeling van burgers, als op het voorkomen van grootschalige privacyinbreuken in hun levens.

4. Verbied biometrische identificatie in de openbare ruimte

Het vierde verbod (artikel5d) gaat over real-time biometrische identificatie op afstand in openbare ruimte voor opsporingsdoeleinden. Een voorbeeld is gezichtsherkenning. De politie mag door dit verbod iemand die op straat loopt niet direct via de camerabeelden identificeren. Het verbod kent drie uitzonderingen. De politie mag de technieken wel inzetten voor het zoeken van slachtoffers van een misdaad en vermiste kinderen, ter voorkoming van een acute dreiging, zoals een terroristische aanslag, en bij het vinden en identificeren van een verdachte van een crimineel vergrijp (de zogeheten ‘Eurocrimes’).

Helaas is er met grote regelmaat een slachtoffer te zoeken, een dader van dergelijke vergrijpen te vinden of een kind vermist. De facto zullen er dus systemen zijn die real-time biometrische identificatie op afstand mogelijk maken. Function creep – het gebruik van het systeem voor andere doelen dan oorspronkelijk vastgelegd – ligt daarbij steeds op de loer. Zo bleek het ANPR, succesvol ingezet bij het lokaliseren van de daders van de aanslag op Peter R. de Vries, ook jarenlang automobilisten en bijrijders te hebben gefotografeerd, terwijl de wettelijke basis hiervoor ontbrak.

Daarmee wordt het steeds moeilijker voor burgers om zich anoniem in de openbare ruimte te begeven. Bovendien werkt de techniek niet vlekkeloos. Systemen kunnen mensen verkeerd identificeren, groepen mensen systematisch benadelen of uitsluiten. Het feit dat je bekeken kan worden, kan leiden tot chilling effecten: mensen kunnen zich uit angst bekeken of geregistreerd te worden anders gaan gedragen. Zo kunnen AI-systemen de vrijheid om samen te komen en te demonstreren inperken. De Europese toezichthouders hebben zich daarom kritisch getoond over de beperkte reikwijdte van het verbod. Ook het Rathenau Instituut pleit voor een breed verbod van biometrische identificatie in de openbare ruimte – niet alleen voor opsporingsdoelen. Ook als burgers elkaar real-time kunnen identificeren in de openbare ruimte, komen mensenrechten in het geding.

De Tweede Kamer kan er bij de minister op aandringen om het verbod op real-time remote biometrische identificatie in de openbare ruimte te verbreden.

5. Vul de lijst van hoog-risico-systemen aan

De verordening geeft ook aan welke AI-systemen een ‘hoog-risico’ kennen, met het oog op gezondheid en veiligheid van mensen en waar grondrechten in het geding komen. In de bijlage noemt de Europese Commissie acht domeinen waarin dit het geval kan zijn: biometrische identificatie, kritieke infrastructuur, onderwijs, toegang tot werk, toegang tot essentiële private of publieke diensten en sociale zekerheid, opsporing, migratie, asiel en grenscontrole, en de rechtspraak. Het Rathenau Instituut ziet in haar onderzoeken in deze domeinen inderdaad een hoog risico op schending van mensenrechten.

Tegelijkertijd ontbreken er toepassingen, bijvoorbeeld in publieke sectoren zoals de zorg, terwijl ook hier gezondheid, veiligheid en grondrechten in het geding kunnen komen. Verder wordt emotieherkenning niet genoemd, terwijl die techniek vergelijkbare risico’s kent als biometrische identificatie. Bovendien is AI nog niet in staat om daadwerkelijk emoties te herkennen. Een gelaatsuitdrukking is niet hetzelfde als een emotie (Barrett et al 2019; Hoegen et al 2019). Tot slot ontbreken deepfakes op de lijst, terwijl ze kunnen leiden tot individuele, organisatorische en maatschappelijke schade. Zo kunnen deepfakes het vertrouwen in wetenschap en journalistiek schaden en de beïnvloeding van politieke keuzes via deepfakes ondermijnt een open en eerlijk democratisch debat en stemproces. Het vertrouwen in de democratische rechtstaat

staat dus op het spel. Gezien deze ernstige risico's verdient het aanbeveling deepfaketechnologie op de lijst van hoog-risico-systemen te plaatsen.

De Tweede Kamer kan de minister vragen de lijst van hoog-risico-systemen aan te vullen met systemen in publieke sectoren, deepfaketechnologie en emotieherkenning.

6. Neem ook de maatschappelijke opbrengst van AI-systemen onder de loep

Voor de systemen met een hoog risico gelden extra eisen voor risicobeheersing. Er moet onder meer een systeem komen voor risicobeheersing, datasets moeten aan kwaliteitscriteria voldoen, de werking van het systeem moet dusdanig transparant zijn dat gebruikers de output van het systeem kunnen begrijpen, en menselijk toezicht is nodig. Dat zijn belangrijke voorschriften.

Maar een inhoudelijke ethische beoordeling vraagt ook om informatie over de effectiviteit van het voorgestelde AI-systeem. Doet het systeem wat het belooft? Lukt het om sociale verzekeringsfraude beter op te sporen of de politie-inzet efficiënter in te zetten? Daar ontbreekt regelmatig het bewijs voor, vond ook het onderzoeksinstituut JRC (2019) in een studie naar Europese overheidssystemen. Ondertussen is de kans op mensenrechtenschending steeds aanwezig: dataverzameling creëert een privacy- en beveiligingsrisico, en profilering creëert kans op vooroordelen, te weinig maatwerk of beperking van autonomie. Het is dus zaak om niet slechts de risico's te wegen, maar ook te kijken naar de maatschappelijke opbrengst van AI.

De Tweede Kamer kan de minister vragen om een betere weging van de maatschappelijke opbrengst, door betere onderbouwing van de effectiviteit van het systeem te eisen.

7. Transparantie is belangrijk, maar niet voldoende

Tot slot komt de AI-verordening met transparantieverplichtingen voor bepaalde type AI-systemen. Het gaat om systemen die specifieke risico's op imitatie of bedrog met zich meebrengen, zoals AI-systemen die met mensen interacteren (een chatbot), emotieherkenning, 'biometrische categorisatie' en AI-systemen waarmee beeld, audio- of videocontent kan worden gemaakt of gemanipuleerd (deepfakes). Bij deze systemen moet steeds duidelijk zijn dat je met een computersysteem praat en niet tegen een mens, en dat het om beelden gaat die kunstmatig gemaakt of gemanipuleerd zijn.

Gezien de ontwikkeling van AI in chatbots, spraakassistenten en deepfakes, is het risico op imitatie en bedrog inderdaad steeds groter. Een transparantieverplichting is belangrijk, maar perkt de risico's onvoldoende in. Ook een gelabelde deepfake of een emotieherkenningssysteem kan leiden tot aanzienlijke individuele en maatschappelijke schade. De melding dat je emoties herkend worden, bijvoorbeeld tijdens een sollicitatie-

of asielprocedure, geeft iemand weinig mogelijkheden om hiertegen protest aan te tekenen. Bovendien kent het voorstel diverse uitzonderingen op de transparantieverplichting. Zo is het niet nodig om deepfakes te labelen voor opsporing, kunst, wetenschap, of waar 'vrijheid van meningsuiting nodig is'. Deze uitzonderingen zijn te breed.

Ook is het belangrijk dat burgers die geschaad worden door AI-beslissingen weten waar ze terecht kunnen, en een mogelijkheid hebben tot genoegdoening. Daar hoort ook een toegankelijke klachten- en bezwaarprocedure bij, waar burgers kunnen melden dat zij op onjuiste of onrechtvaardige manier worden behandeld.

De Tweede Kamer kan bij de minister aandringen op effectieve rechtsbescherming voor benadeelde burgers, en vragen de uitzonderingen op transparantie te beperken.

Relevante publicaties van het Rathenau Instituut

Grip op algoritmische besluitvorming bij de overheid. De rol van de Eerste Kamer (Notitie ter ondersteuning van de Werkgroep AI, 2021)

Stel nu 10 ontwerpeisen aan de digitale samenleving van morgen (Rathenau Manifest, 2020)

Tackling deepfakes in European policy (Rapport op verzoek van het Panel for the Future of Science and Technology (STOA), Europees Parlement, 2021).

Zeven acties voor verantwoord innoveren met AI (Bericht aan het parlement, 2020)

Investeer in digitalisering die werkt voor mensen (Bericht aan het parlement, 2020)

De toekomst van online platformen (Bericht aan het parlement, 2020)
