Hoofdstuk 4. Casestudy Contentmoderatie door online platformen

Stefan Kulk & Thom Snijders

4.1 Introductie

Het internet biedt talloze nieuwe mogelijkheden waardoor individuen online met elkaar kunnen interacteren en communiceren. Online dienstverleners zoals Facebook, Youtube en Twitter bieden mensen een platform om zichzelf te uiten en informatie met elkaar te delen. De platformen die worden geboden kunnen echter ook gebruikt worden voor het verspreiden van content die om tal van redenen als ongewenst kan worden beschouwd door mensen, de dienstverlener en de maatschappij in brede zin. Te denken valt aan het verspreiden van desinformatie, het doen van discriminerende uitingen en het delen van terroristische propaganda. Maar bijvoorbeeld ook de verspreiding van naaktfoto's kan door bepaalde platformen als ongewenst worden bestempeld.

Online platformen kunnen een belangrijke rol spelen in het voorkomen en stoppen van de verspreiding van dergelijke content. Zij stellen regels op en nemen beslissingen over de wijze waarop bepaalde content wordt weergegeven en bepalen of content minder prominent zichtbaar moet zijn, verwijderd dient te worden of überhaupt online komt. Dit proces wordt ook wel contentmoderatie genoemd.[1] Voor het modereren van content wordt door platformen in toenemende mate gebruik gemaakt van algoritmen.

Het modereren van online content door online platformen is het onderwerp van deze casestudy. Daarin schenken we in het bijzonder aandacht aan het modereren van *hate speech*, waaronder ook discriminerende uitingen. Omdat het vaststellen van de onrechtmatigheid van dergelijke content afhankelijk is van feiten en omstandigheden, en algoritmen maar in beperkte mate rekenschap kunnen geven van de context waarin een uiting wordt gedaan, kan met een nadere bestudering van het modereren van *hate speech* worden blootgelegd wat de mogelijkheden en onmogelijkheden zijn van algoritmen in dit domein.

4.1.1 Methodologie

Het is allereerst van belang om op te merken dat het slechts beperkt mogelijk is om een volledig beeld te schetsen van de wijze waarop online platformen content modereren. De technologieën die platformen inzetten om content te modereren zijn in private handen. Platformen zijn in beginsel

[1] Contentmoderatie is geen nieuw fenomeen. De noodzaak van het maken van keuzes met betrekking tot welke content wordt getoond aan gebruikers of consumenten is ouder dan de hedendaagse online platformen, en in zekere zin zelfs ouder dan het internet; kranten, televisiezenders en andere traditionele media hebben allemaal te maken met het cureren van hun aanbod. Zie Gillespie 2018, p. 74. Contentmoderatie wordt in deze casestudy in meer detail uitgelegd in par. 4.1.3; voor een wetenschappelijke definitie van *content moderation*, zie bijvoorbeeld Roberts 2019.

niet verplicht om openheid van zaken te geven met betrekking tot de werking van hun technologieën en de inzet van algoritmen om content te modereren. Als het gaat om de werking en toepassing van algoritmen voor contentmoderatie baseren we ons daarom enerzijds op de publieke (journalistieke) informatie over het modereren van content door platformen en anderzijds op wetenschappelijk onderzoek naar de wijze waarop algoritmen ingezet *kunnen* worden.

Om vergaarde informatie aan te vullen en verkregen inzichten te verifiëren, zijn semigestructureerde interviews afgenomen. Meer in het bijzonder was het doel van de interviews om een beter beeld te krijgen van de werking en het gebruik van contentmodereeralgoritmen, evenals de context waarin ze worden gebruikt. Daarnaast is verkregen input gebruikt ten behoeve van de inventarisatie van de kansen en risico's die zich kunnen voordoen met betrekking tot de geïdentificeerde publieke waarden en de houdbaarheid van het juridisch kader.[2]

4.1.2 Opzet van de casestudy

In deze casestudy bespreken we eerst, in algemene zin, hoe algoritmen kunnen worden gebruikt voor het modereren van onrechtmatige en onwenselijke online content (par. 4.1.3). Vervolgens spitsen we de casestudy toe op het fenomeen *hate speech* en bespreken we de algoritmen die platformen inzetten om de verspreiding van *hate speech* tegen te gaan (par. 4.2). Daarna worden per publieke waarde de kansen en risico's van het gebruik van dergelijke algoritmen geïnventariseerd en wordt het relevante juridisch kader in kaart gebracht en geëvalueerd (par. 4.3 t/m 4.6). In tegenstelling tot andere casestudy's bespreken wij daar niet eerst de casestudy-overstijgende waarden, maar vangt onze analyse aan met een bespreking van de casestudyspecifieke publieke waarde van vrijheid van meningsuiting. Wij hebben daarvoor gekozen omdat het modereren van online content in de kern vooral deze publieke waarde raakt.

4.1.3. Contentmoderatie door algoritmen

Een scherp afgebakende definitie van 'online content' is moeilijk te geven. Wij verstaan daaronder iedere tekstuele, visuele of hoorbare inhoud die online beschikbaar wordt gesteld. Te denken valt aan berichten, foto's, muziek en alle andere inhoud die door gebruikers op online platformen worden geplaatst. Ook reacties op nieuwsberichten, Facebook- of Instagramposts of tweets zijn 'content'. Advertenties die op platformen worden geplaatst kunnen ook als content worden aangemerkt. Hieronder bespreken we wanneer dergelijke content wordt aangemerkt als onrechtmatig of onwenselijk, leggen we uit wat contentmoderatie inhoudt en gaan we in op de wijze waarop algoritmen daarbij ingezet kunnen worden.

[2] In deze casestudy wordt daarom niet verwezen naar de interviews en de daarin verkregen informatie en door de geïnterviewden naar voren gebrachte zienswijzen.

4.1.3.1 Onrechtmatige en onwenselijke online content

Er is een breed scala aan typen onrechtmatige en onwenselijke online content die met behulp van algoritmen kunnen worden gemodereerd. Zij kunnen langs tenminste drie lijnen worden onderscheiden.

In de eerste plaats dient er een onderscheid gemaakt te worden tussen onrechtmatige content en ongewenste content. In het geval van ongewenste content is er geen sprake van overtreding van een juridische norm door het maken of verspreiden van de content. Te denken valt aan online desinformatie, waarvan de creatie en verspreiding een verstorend effect kan hebben op bijvoorbeeld democratische processen, maar die niet per definitie onrechtmatig is.[3] Ook kan worden gedacht aan content die door online platformen als ongewenst wordt aangemerkt en die verwijderd kan worden op grond van de beleidsregels van het platform, maar die niet onrechtmatig is, zoals naaktbeelden.[4] Het zijn dan de opvattingen en denkbeelden van de platformbeheerder die de doorslag geven.

In de tweede plaats kan er ten aanzien van onrechtmatige content een onderscheid worden gemaakt naar het karakter van de onrechtmatigheid. Zo is er content die naar zijn aard onrechtmatig is. Voorbeelden zijn kinderpornografisch materiaal en content waarin aangezet wordt tot haat of geweld, of waarin een groep wordt beledigd.[5] Daarnaast is er een categorie van content die op zichzelf niet onrechtmatig is, maar waarbij de onrechtmatigheid is gelegen in de verspreiding ervan en de voorwaarden waaronder dat gebeurt. Voorbeelden zijn de ongeautoriseerde openbaarmaking van auteursrechtelijk beschermd werk of privacyinbreuken. De content in kwestie (bijvoorbeeld een film of een muziekstuk, of een nietsverhullende foto) zelf is dan niet onrechtmatig, maar de handeling om die zonder toestemming openbaar te maken wel.[6]

In de derde plaats kan er onderscheid gemaakt worden naargelang het effect dat de content heeft. Een inbreuk op een intellectueel eigendomsrecht zal veelal economische schade veroorzaken.[7]

---

[3] De High-Level Expert Group on fake news and online disinformation zegt hierover het volgende: '*Disinformation as defined here includes forms of speech that fall outside already illegal forms of speech, notably defamation, hate speech, incitement to violence, etc. but can nonetheless be harmful*.' HLEG on Fake News and Disinformation 2018, p. 10. Zie in dat verband ook de uitspraak van het EHRM in de zaak *Salov*, waarin is bepaald dat ook het verspreiden van informatie, waarvan een sterk vermoeden bestaat dat die niet waar is, onder de bescherming van art. 10 EVRM valt: EHRM 6 december 2005, ECLI:CE:ECHR:2005:0906JUD006551801 (*Salov/Oekraïne*).

[4] Het delen van dergelijke content zou echter wel in strijd kunnen zijn met contractuele voorwaarden waaraan gebruikers van platformen zich hebben verbonden.

[5] Art. 240b Sr verbiedt bijvoorbeeld de vervaardiging en verspreiding van kinderpornografisch materiaal. Art. 137d Sr verbiedt het aanzetten tot haat of geweld en art. 137c Sr stelt groepsbelediging strafbaar. Daarnaast kan in het kader van contentmoderatie ook het strafrechtelijke verbod van art. 137e Sr relevant zijn, dat ziet op de openbaarmaking van uitlatingen in de zin van art. 137c juncto art. 137d Sr, en art. 137f Sr, dat onder andere deelname aan een activiteit gericht op discriminatie strafbaar stelt.

[6] Hoewel er ten aanzien van dit soort content strikt genomen geen sprake is van onrechtmatige content, spreken we uit overwegingen van de leesbaarheid van deze casestudy wel van onrechtmatige content.

[7] Persoonlijkheidsrechten kunnen echter wel een rol spelen.

Voor content zoals 'wraakporno'[8] of discriminerende uitingen is vooral de menselijke waardigheid in het geding en zullen economische overwegingen een minder grote rol spelen. Een tussenvorm is wellicht de schending van portretrechten waarbij zowel economische belangen als privacybelangen een rol kunnen spelen.[9] Daarnaast zijn er typen content die de samenleving als geheel kunnen raken en waarvan de verspreiding dus om die reden onwenselijk wordt geacht. Daartoe behoren bijvoorbeeld terroristische inhoud of online desinformatie, maar ook discriminerende uitingen.

4.1.3.2 Contentmoderatie

Ten aanzien van content die is aangemerkt als onrechtmatig, ligt het verwijderen daarvan voor de hand. Als een rechter heeft bepaald dat, bijvoorbeeld, een post op Facebook onrechtmatig is, dan dient die post verwijderd te worden. Contentmoderatie is echter een proces dat doorgaans geheel plaatsvindt bij het platform zelf. Het platform ontvangt meldingen over, of gaat zelf op zoek naar, onrechtmatige of onwenselijke content. Het platform beoordeelt de aangebrachte of gevonden content zelf, en neemt vervolgens actie waar nodig. Van rechterlijke tussenkomst is zeer zelden sprake.

De laatste jaren is er vanuit verschillende hoeken aandacht voor de verantwoordelijkheid van online platformen om de verspreiding van onrechtmatige en ongewenste content op hun platformen tegen te gaan.[10] Het kan daarbij gaan om het (snel) verwijderen van content, maar ook het deprioriteren (lager *ranken*) daarvan.[11] Beheerders van platformen kunnen er ook zelf belang bij hebben om bepaalde vormen van onrechtmatige of onwenselijke content te weren, bijvoorbeeld om gebruikers en adverteerders aan zich gebonden te houden.[12] Om onrechtmatige en onwenselijke content te vinden, wordt tegenwoordig lang niet meer alleen gebruik gemaakt van

[8] Het begrip 'wraakporno' is een misleidende term omdat het eigenlijk gaat om de ongeautoriseerde verspreiding van beelden van seksuele aard. Wraak hoeft daarin geen rol te spelen. Bovendien wordt door het begrip 'wraak' gesuggereerd dan het verspreiden van de beelden in zekere zin te billijken zou zijn. Meer daarover Sebastian, *Feminist Media Studies* 2017, p. 1107.

[9] Zie voor een recent voorbeeld daarvan in de Nederlandse rechtspraak: Rb. 11 november 2019, ECLI:NL:RBAMS:2019:8415 (*John de Mol/Facebook*), over de 'nepadvertenties' voor cryptovaluta waarin beelden van John de Mol werden gebruikt.

[10] Zie bijvoorbeeld Chesney & Citron, *Foreign Affairs* 2019, p. 147.

[11] Het ranken van informatie als zodanig valt buiten het bereik van deze casestudy. Zie daarover onder meer de Platform-to-business Verordening waarin in art. 5 een verplichting tot verstrekking van informatie over het ranken van informatie door zoekmachine en e-commerce platformen is opgenomen (Verordening (EU) 2019/1150 van het Europees Parlement en de Raad van 20 juni 2019 ter bevordering van billijkheid en transparantie voor zakelijke gebruikers van onlinetussenhandelsdiensten (Voor de EER relevante tekst), *PbEU* 2019, L 186). Zie ook het werk van het Observatory on the Online Platform Economy (platformobservatory.eu). Zie ook het burgerinitiatief 'internetpesters aangepakt' en de Kamerbrief daarover van 17 juli 2019 (*Kamerstukken II* 2018/19, 34602, nr. 2).

[12] Bijvoorbeeld Van Noort, *NRC* 13 februari 2018; Dang, *Reuters* 21 februari 2019; Harding, *CBS NEWS* 11 mei 2018; 'Mars pulls ads from YouTube drill videos', BBC.com 4 augustus 2018; zie met betrekking tot Dumpert, het videoplatform van GeenStijl: 'Adverteerders stoppen met adverteren op Dumpert en GeenStijl', *NOS*.nl 3 mei 2017.

meldingen van mensen.[13] Om proactief te kunnen optreden zetten platformen ook algoritmen in om deze typen content te detecteren zodra gebruikers die online beschikbaar willen stellen.[14]

In zowel het privaatrecht als het publiekrecht zijn er normen die ertoe strekken dat platformen bepaalde onrechtmatige content verwijderen en, afhankelijk van de content in kwestie, ook zelf opsporen. Het aansprakelijkheidsrecht, met name de aansprakelijkheidsbeperkingen in art. 6:196c BW, stimuleren platformen om onrechtmatige content te verwijderen zodra zij daarvan kennis hebben.[15] Maar ook in het publiekrecht zien we terug dat van platformen wordt verlangd dat zij content modereren. Een voorbeeld is het conceptvoorstel voor een EU-verordening ter voorkoming van de verspreiding van terroristische online-inhoud, dat platformen verplicht terroristische content binnen een uur na melding van een autoriteit te verwijderen en verwijderd te houden.[16]

Platformen kunnen ook zelf op grond van een gebruikersovereenkomst (*terms of service*) paal en perk stellen aan de verspreiding van onwenselijke of onrechtmatige content op het platform. Deze overeenkomsten stellen online platformen in staat om content van gebruikers te verwijderen of de toegang daartoe te beperken als die de huisregels van het platform schenden.[17] De typen content die door platformen niet zijn toegestaan komen in grote lijnen overeen met de typen content die ook juridisch gezien onrechtmatig zijn. Als private ondernemingen staat het platformen echter vrij om ook strengere regels te hanteren en content die op zichzelf niet onrechtmatig is, zoals naaktbeelden, van het platform te weren.

4.1.3.3 Inzet van algoritmen

Verschillende typen onrechtmatige en ongewenste content vragen om een eigen aanpak. De rol die platformen daarin (kunnen) spelen en de wijze waarop algoritmen worden ingezet, verschilt daarom ook per type content.

Algoritmen kunnen een rol spelen bij het herkennen van onrechtmatige of ongewenste content. Zo kan een algoritme door tekstanalyse herkennen of er in een bepaald geval mogelijk sprake is van het aanzetten tot haat, geweld of discriminatie. Die content kan dan worden doorgeleid naar een medewerker van het platform, die er een definitief oordeel over velt. Het is ook denkbaar dat een algoritme detecteert dat een gebruiker onrechtmatige content probeert te delen en zelf besluit de

[13] Ook partijen die klagen over content bij platformen kunnen zelf algoritmen gebruiken. Zo is bekend dat rechthebbenden algoritmen inzetten voor het vinden van auteursrechtinbreuken op internet (Urban, Karaganis & Schofield 2016). Een ander voorbeeld is de deels geautomatiseerde detectie van *hate speech* door 'Hatebusters'. Dit programma legt gebruikers reacties op YouTube voor die mogelijk *hate speech* bevatten zodat zij die reacties kunnen aanbrengen bij Youtube (hatebusters.org).
[14] Gillespie 2018, p. 77.
[15] Deze bepaling implementeert artt. 12-14 van de E-Commercerichtlijn (Richtlijn 2000/31/EG van het Europees Parlement en de Raad van 8 juni 2000 betreffende bepaalde juridische aspecten van de diensten van de informatiemaatschappij, met name de elektronische handel, in de interne markt ("Richtlijn inzake elektronische handel"), *PbEG* 2000, L 178).
[16] Zie voor het wetgevingsdossier met betrekking tot deze verordening 2018/0331(COD).
[17] Zie voor een voorbeeld van een dergelijke contractuele bepaling punt 2 van de Facebook servicevoorwaarden (perma.cc/3H7P-BQXS). Voor een voorbeeld van dergelijke huisregels zie facebook.com/communitystandards.

content te weren, zoals reeds gebeurt met betrekking tot auteursrechtelijk beschermde werken.[18] Op basis van een *hash* - een digitale vingerafdruk die wordt gemaakt van een beschermd werk - kan dat werk door een algoritme worden herkend en kan worden voorkomen dat bepaalde content online verschijnt. Een gevaar is dat dergelijke algoritmen in geüpload materiaal ten onrechte een auteursrechtelijk beschermd werk herkennen.[19] Maar ook als materiaal correct wordt herkend, is nog niet gezegd dat dat materiaal ook daadwerkelijk inbreukmakend is. Zo staat het auteursrecht gebruik van werken toe als er bijvoorbeeld sprake is van een parodie of een citaat. Of daarvan sprake is, vergt soms een complexe juridische afweging, die (vooralsnog) niet door het algoritme gemaakt kan worden.

Algoritmen die *hashes* vergelijken kunnen ook worden ingezet om te voorkomen dat reeds als onrechtmatig aangemerkte content zich verder verspreidt. Te denken valt aan kinderpornografie of terroristische content. Ook van dergelijke content kan een *hash* gemaakt worden die door een algoritme kan worden vergeleken met nieuw geüploade content.[20]

Andere algoritmen worden ingezet om de authenticiteit van content vast te stellen en worden zodoende gebruikt om bepaalde vormen van online desinformatie te detecteren. Een voorbeeld van dergelijke content betreft zogeheten '*deep fakes*': gemanipuleerde video's die (bijna) niet van echt zijn te onderscheiden en het publiek in verwarring kunnen brengen.[21] Voorbeelden zijn video's waarin wereldleiders woorden in de mond gelegd worden die zij niet hebben gezegd.[22] Zelflerende algoritmen kunnen worden ingezet om op basis van gelabelde datasets van video's, gemanipuleerde video's te leren herkennen.[23] Een andere methode kan zijn om video's op bepaalde veelvoorkomende eigenschappen van *deep fakes*, zoals de lage resolutie van gemanipuleerde beelden, met behulp van neurale netwerken te ontdekken.[24] Met betrekking tot desinformatie kunnen algoritmen ook worden ingezet om links naar gefactcheckte artikelen te vinden, zodat automatisch een factchecklabel bij berichten kan worden geplaatst.[25]

[18] Google heeft daarvoor 'ContentID' ontwikkeld. Het systeem van Facebook heet 'Rights Manager'. Beide systemen checken geüpload materiaal tegen een database van beschermde werken.
[19] Lester & Pachamanova, *UCLA Entertainment Law Review* 2017, p. 51-73. Zie daarover ook Kulk 2019, p. 280.
[20] Zie bijvoorbeeld gifct.org/joint-tech-innovation/, waarin Facebook, Microsoft, Twitter, and YouTube samenwerken om de verspreiding van terroristische content te voorkomen en een '*database of hashes*' onderhouden waarmee wordt voorkomen dat bepaalde content opnieuw op een van de platformen verschijnt. Zie ook 'Open-Sourcing Photo- and Video-Matching Technology to Make the Internet Safer', fb.com 1 augustus 2019, waarin Facebook open source software beschikbaar stelt om identieke of bijna identieke beelden te herkennen.
[21] Voorbeelden zijn video's waarin het gezicht of alleen de mond en lippen van iemand zijn vervangen. De term is een porte-manteau van de termen '*deep learning*' en '*fake*'.
[22] Zie over de opkomst en gevaren van *deep fakes* in geopolitieke context: Citron & Norton, *Boston University Law Review* 2011, p. 1435.
[23] Zie voor een dergelijke dataset: Deepfake Detection Challenge, deepfakedetectionchallenge.ai. Zie over deze dataset: Dolhansky e.a. 2019. Zie ook over de FaceForensics Dataset: Rössler e.a. 2018.
[24] Zie bijvoorbeeld Li & Lyu 2019. Voor een overzicht van verschillende methoden om *deep fakes* te herkennen, zie Nguyen e.a. 2019.
[25] Zie over zulke labels bijvoorbeeld Gibbs, *The Guardian* 7 april 2017.

In alle gevallen is het proportionaliteitsvereiste van belang voor de invulling van contentmoderatie en de rol die algoritmen daarin kunnen spelen. De aanpak die wordt gekozen, moet in verhouding staan tot onder andere de effecten van de content in kwestie en de gevolgen van de inzet van algoritmen. Daarnaast is van belang dat het modereren van content slechts één middel is om de verspreiding van onrechtmatige en ongewenste content tegen te gaan. Het bestrijden van online desinformatie is daarvan een goed voorbeeld. De aanpak van online desinformatie vergt een combinatie van maatregelen, zoals het bevorderen van mediawijsheid, de inzet van *factcheckers*, en het verzekeren van de pluriformiteit van de media.[26] Ook discriminerende uitingen kunnen dichter bij de bron worden voorkomen, bijvoorbeeld door middel van educatieve campagnes.[27]

Het modereren van content door middel van algoritmen hoeft bovendien niet altijd te leiden tot het verwijderen van content. Om de negatieve effecten van bijvoorbeeld online desinformatie te mitigeren, zetten platformen algoritmen in om de betrouwbaarheid van bepaalde content te wegen en wordt onbetrouwbare informatie een lagere positie gegeven in de ordening van de content (*ranking*).[28] Het modereren van online content is dus meer dan alleen het verwijderen van onrechtmatige of onwenselijke content.

4.2 De aanpak van *hate speech* door online platformen

In deze casestudy gaan we nader in op de inzet van algoritmen voor de aanpak van online *hate speech*. Daarvoor zal allereerst het fenomeen van online *hate speech* kort worden besproken. Daarna bespreken we de werking en toepassing van de verschillende soorten algoritmen die worden ingezet om *hate speech* te detecteren. Ook gaan we in op hoe het gebruik van algoritmen zich in dit domein in de toekomst zou kunnen ontwikkelen.

4.2.1 Hate speech

De term *hate speech*, die oorspronkelijk uit de Verenigde Staten afkomstig is, kan het best worden beschouwd als een verzamelbegrip dat wordt gebruikt om (onderdelen van) een spectrum van schadelijke of anderszins onwenselijke uitingen aan te duiden. Daaronder vallen het oproepen tot geweld, haatdragende en haatzaaiende uitingen, maar mogelijk ook andere zeer beledigende uitingen en uitingen die getuigen van extreme vooroordelen en/of vooringenomenheid.[29] Het gaat daarbij onder andere om openbare op schrift gestelde of door middel van afbeelding gedane uitingen. Over zowel de precieze reikwijdte van de term *hate speech* als het onderscheidende

[26] HLEG on Fake News and Disinformation 2018, p. 35. Zie ook McGonagle e.a. 2018.
[27] Zie Titley, Keen & Földi 2014, p. 41.
[28] Bijvoorbeeld Google, over de inzet van algoritmen voor het herkennen van desinformatie ten aanzien van zijn producten 'How Google Fights Disinformation', Google februari 2019, blog.google/documents/37/How_Google_Fights_Disinformation.pdf, p. 4.
[29] McGonagle 2013, p. 4. Over de oorsprong van de term *hate speech*, zie Brown, *Law and Philosophy* 2017, p. 424.

karakter ervan bestaan uiteenlopende opvattingen.[30] Volgens sommigen kenmerkt *hate speech* zich in het bijzonder door een intentie om de (gelijk)waardigheid van bepaalde groepen personen - en daarmee de maatschappelijke acceptatie - te ondermijnen.[31] Voor anderen is vooral het extreme karakter van de uiting van centraal belang.[32]

Hoewel *hate speech* niet is voorbehouden aan het online domein, is dit een ruimte waarin *hate speech* zich makkelijker, sneller en verder kan verspreiden dan in de offline wereld, en die gebruikers daarbij ogenschijnlijk een hoge mate van anonimiteit biedt.[33] Online *hate speech* kan bovendien een katalysator zijn voor offline geweld, en de impact ervan reikt vaak verder dan de persoon waartegen een specifieke uiting is gericht.[34] Nagenoeg alle grote online platformen hebben regels met betrekking tot *hate speech* en verbieden daarbij in ieder geval het oproepen tot geweld en haatdragende en haatzaaiende uitingen; hierbij wordt veelal geput uit juridische terminologie uit bijvoorbeeld Amerikaanse antidiscriminatiewetgeving, met name ten aanzien van beschermde of kwetsbare groepen.[35] De in deze regels gehanteerde definities van *hate speech*, en de soorten uitlatingen die als voorbeeld worden gegeven van wat als ontoelaatbaar wordt beschouwd, verschillen desalniettemin per platform. Sommige platformen geven er de voorkeur aan zich niet te snel te mengen in het online debat en slechts de grootste uitwassen, zoals het oproepen tot geweld, te bestrijden, terwijl andere platformen strengere regels hanteren en proactiever optreden.[36]

Er is op zowel Europees als nationaal niveau aandacht voor de bestrijding van (online) *hate speech*. Zo heeft de Raad van Europa diverse aanbevelingen gedaan om *hate speech* te bestrijden.[37] In de EU is met name het Kaderbesluit racisme en vreemdelingenhaat van belang.[38] Dit Kaderbesluit vormt ook de basis voor de *Code of Conduct on Countering Illegal Hate Speech Online* die later in deze casestudy besproken zal worden.[39] Daarnaast verplicht de Richtlijn

[30] Zie voor verschillende definities o.a. Rosenfeld, *Cardozo Law Review* 2003, p. 1523; Cohen-Almagor, *Policy & Internet* 2011, p. 1.
[31] Waldron 2012, p. 5.
[32] Post 2009, p. 123.
[33] Zie Wilson 2012, p. 3; López & López 2017, p. 11-12. Zie met betrekking tot de rol van anonimiteit Mondal, Silva & Benevenuto 2017.
[34] Een bekend voorbeeld van hoe online *hate speech* fysiek geweld kan aanwakkeren is de bewuste inzet van *hate speech* in Myanmar. Zie Mozur, *The New York Times* 15 oktober 2018; Wilson 2012, p. 4. Zie ook Müller & Schwarz 2018.
[35] Gillespie 2018, p. 58. Zie voor het beleid van Facebook nl-nl.facebook.com/communitystandards/hate_speech. Zie voor het beleid van Twitter: help.twitter.com/nl/rules-and-policies/hateful-conduct-policy. Zie voor het beleid van YouTube: support.google.com/youtube/answer/2801939?hl=nl.
[36] Zie in dat verband bijvoorbeeld de aanscherping van het beleid dat Twitter voert (Conger, *The New York Times* 9 juli 2019). En met betrekking tot het online lastigvallen van mensen, zie Pater e.a. 2016, p. 369.
[37] Zie bijvoorbeeld Aanbeveling (97) 20 van het Comité van Ministers van de Raad van Europa (30 oktober 1997), *On hate speech*. Voor een overzicht van de activiteiten van de Raad van Europa op dit gebied, zie 'Freedom of expression. Hate speech', coe.int. Daarnaast monitort de *European Commission against Racism and Intolerance* (ECRI) het bestaan van racisme en intolerantie in Europa, zie daarvoor 'European Commission against Racism and Intolerance (ECRI)', coe.int. En zie met betrekking tot *hate speech* ECRI General Policy Recommendation no. 15 (8 december 2015) *On Combating Hate Speech*.
[38] Kaderbesluit 2008/913/JBZ van de Raad van 28 november 2008 betreffende de bestrijding van bepaalde vormen en uitingen van racisme en vreemdelingenhaat door middel van het strafrecht (*PbEU* 2008, L 328).
[39] Zie par. 4.3.1 voor een nadere bespreking. *Code of conduct on countering illegal hate speech online*, 30 juni 2016, ec.europa.eu/newsroom/just/document.cfm?doc_id=42985.

Audiovisuele mediadiensten lidstaten om ervoor te zorgen dat videoplatformen passende maatregelen nemen om het publiek te beschermen tegen video's die aanzetten tot geweld of haat.[40]

In Nederland zijn de belangrijkste juridische instrumenten ten aanzien van de bestrijding van *hate speech* het strafrechtelijke verbod op groepsbelediging en het verbod op haatzaaien.[41] Ook het verbod op de openbaarmaking van zulke uitingen en het verbod op o.a. deelname aan activiteiten die gericht zijn op discriminatie zijn in het kader van online *hate speech* relevant.[42] Ten aanzien van de aanpak van online *hate speech* speelt in Nederland ook het Meldpunt Internetdiscriminatie (MiND) een rol.[43] Personen kunnen discriminerende uitingen melden bij MiND, dat vervolgens een inschatting maakt van de strafbaarheid van de uiting in kwestie en een platform kan verzoeken de uiting te verwijderen. Als een verwijderverzoek niet wordt opgevolgd, kan dat ook leiden tot een melding aan het Openbaar Ministerie.

4.2.2 De werking van contentmodereeralgoritmen

Als het gaat om het domein van *hate speech*, dan worden algoritmen primair door platformen gebruikt om uitingen te detecteren die mogelijk kwalificeren als *hate speech*. Deze uitingen worden dan ter beoordeling voorgelegd aan een menselijke moderator. *Hate speech* kan vele vormen aannemen. Het hoeft bij *hate speech* niet alleen maar te gaan om geschreven tekst, maar er kan ook sprake zijn van (een combinatie van) afbeeldingen, audio en video's. Te denken valt aan zogeheten internetmemes, waarin tekst bijvoorbeeld wordt gecombineerd met een sprekende afbeelding. Het geheel van tekst en afbeelding kan dan gelden als *hate speech*. Het zijn met name beelden, en teksten die zijn opgenomen *in* een afbeelding of video, die voor algoritmen moeilijk te herkennen zijn.[44]

Er bestaat een breed spectrum van technologieën die kunnen worden ingezet voor het detecteren van *hate speech*. Een betrekkelijk eenvoudige manier om *hate speech* te detecteren is het gebruik van woordfilters. Er wordt dan door software gecheckt of er sprake is van gebruik van een woord op basis van een zwarte lijst van 'verboden' woorden.[45] In feite gaat het dan om een regelgebaseerd algoritme dat aanslaat op gebruik van vooraf bepaalde woorden. Eenvoudige woordfilters doen echter geen recht aan de context waarin het woord of de combinatie van woorden

[40] Richtlijn 2018/1808 van het Europees Parlement en de Raad van 14 november 2018 tot wijziging van Richtlijn 2010/13/EU betreffende de coördinatie van bepaalde wettelijke en bestuursrechtelijke bepalingen in de lidstaten inzake het aanbieden van audiovisuele mediadiensten (richtlijn audiovisuele mediadiensten) in het licht van een veranderende marktsituatie (*PbEU* 2018, L 303). Zie over die verplichting art. 28 ter lid 1 onder b van de richtlijn.
[41] Artt. 137c en 137d Sr.
[42] Respectievelijk artt. 137e en 137f Sr.
[43] Zie mindnederland.nl.
[44] Maar zie ook Sivakumar & Gordo, Paluri, engineering.fb.com 11 september 2018; 'Advancing self-supervision, CV, NLP to keep our platforms safe', ai.facebook.com 1 mei 2019.
[45] Instagram heeft een functionaliteit aan gebruikers aangeboden om zelf een (aanvullende) lijst van verboden woorden te bepalen. 'Keeping Comments Safe on Instagram', instagram.tumblr.com/post/150312324357/160912-news/embed.

wordt gebruikt en is er een grote kans dat uitingen door het algoritme ten onrechte als mogelijke *hate speech* worden gekwalificeerd, of dat het algoritme bepaalde vormen van *hate speech* juist níet aanbrengt.[46] Dergelijke algoritmen zijn namelijk niet goed in staat om de daadwerkelijke betekenis van een uiting, die voor een kwalificatie als *hate speech* van groot belang is, goed te interpreteren. Uitingen die minder expliciet zijn, of die sarcasme of ironie bevatten, zijn moeilijk te herkennen voor computersystemen.[47] Tegelijkertijd hoeft het gebruik van scheldwoorden niet vanzelfsprekend een indicatie te zijn van *hate speech*. Daarnaast kunnen bijvoorbeeld woorden die in principe een niet-pejoratieve betekenis hebben ook als scheldwoord worden gebruikt (denk aan '*gay*' of 'homo').[48] Bovendien bestaat er een kans dat gebruikers algoritmen misleiden door bijvoorbeeld woorden opzettelijk verkeerd te spellen of woorden met een positieve connotatie toe te voegen.[49]

Online platformen gebruiken dan ook steeds vaker zelflerende algoritmen die zij zelf ontwikkelen of die worden aangeboden door derde partijen.[50] Bij vormen van *supervised machine learning* worden uitingen van *hate speech* eerst handmatig als zodanig gelabeld. Die data worden gevoed aan het algoritme opdat het patronen gaat herkennen en deze 'kennis' kan toepassen bij het beoordelen van andere toekomstige uitingen. Voor het creëren van de benodigde datasets kunnen platformen de eerdere beslissingen van menselijke contentmoderators gebruiken, maar het labelen van de data kan ook worden uitbesteed via platformen als Amazon's *Mechanical Turk* of *CrowdFlower*, waar derden de data labelen tegen een kleine vergoeding.[51]

In geval van *supervised machine learning* is de juistheid van een beslissing sterk afhankelijk van de kwaliteit van de (gelabelde) data die wordt gebruikt in trainingsproces. In dat verband is het van belang dat data op consistente wijze worden gelabeld, door mensen met voldoende inhoudelijke kennis.[52] Een probleem kan zijn dat de datasets van uitingen die worden gebruikt om algoritmen te trainen, niet representatief zijn voor het type content dat wordt gemodereerd. Als ook gebruik wordt gemaakt van informatie over gebruikers, dan ligt het gevaar van een bevooroordeeld algoritme op de loer. Daarvan kan sprake zijn als de dataset waarop is getraind een onvolledig beeld schetst van de *hate speech* postende gebruiker.[53] Veranderende interne regels van het platform ten aanzien van wat precies mag en wat niet, kunnen de consistentie van de dataset – en daarmee de voorspelbaarheid van de uitkomst – eveneens negatief beïnvloeden.

[46] Warner & Hirschberg 2012; Davidson e.a. 2017; MacAvaney e.a., *PLoS ONE* 2019.
[47] Pavlopoulos, Malakasiotis & Androutsopoulos 2017, p. 1125.
[48] Davidson e.a. 2017.
[49] Gröndahl e.a. 2018.
[50] Voorbeelden van derde partijen zijn: Utopia AI Moderator (utopiaanalytics.com/utopia-ai-moderator/) en Hatebase (hatebase.org/).
[51] Matsakis, WIRED.com 22 maart 2018.
[52] Waseem 2016.
[53] MacAvaney e.a., *PLoS ONE* 2019.

Een daaraan gerelateerd probleem is dat wat geldt als *hate speech* en de wijze waarop mensen zich uitdrukken na verloop van tijd kan veranderen. Als algoritmen niet worden doorontwikkeld, kan dat leiden tot een verminderde nauwkeurigheid en dus een grotere kans op fout-positieve en fout-negatieve resultaten. Daarnaast kunnen opvattingen over wat *hate speech* is verschillen per taal en cultuur. Dat betekent dat wanneer online platformen bij de ontwikkeling van algoritmen vertrekken vanuit een specifiek cultuurgebonden begrip van *hate speech*, dit nadelige gevolgen kan hebben voor de nauwkeurigheid waarmee *hate speech* in andere culturen en talen wordt herkend.[54]

Natural language processing-technologie speelt een grote rol in het vinden van geschreven *hate speech*, met name als het gaat om het beoordelen van de inhoud van de uiting. *Natural language processing* is een subdomein binnen het domein van kunstmatige intelligentie dat zich bezighoudt met geschreven taal. Binnen dit subdomein wordt onder andere gebruik gemaakt van zelflerende algoritmen die in datasets patronen kunnen ontdekken met betrekking tot de zinsstructuur en inhoud van uitingen. Voorbeelden zijn aspecten zoals het sentiment van een tekst of tekstdeel.[55] Zulke elementen worden betrokken in het oordeel dat een modereeralgoritme vervolgens velt over een uiting.[56]

Ook andere gegevens over zowel de uiting als de gebruiker kunnen worden betrokken in het beoordelingsproces.[57] Gegevens over individuele gebruikers, zoals een online *track record* met daarin bijvoorbeeld informatie over of de gebruiker eerder is berispt voor het schenden van huisregels, kunnen worden gecombineerd met de resultaten van de bovengenoemde inhoudelijke analyse.[58] Andere gegevens over een specifieke uiting, zoals de lengte van de uiting en de mate waarin de inhoud is gerelateerd aan de inhoud van de originele post waaronder deze is geplaatst, kunnen eveneens als indicatoren worden gebruikt. Ook de kans dat bepaalde content *hate speech* ontlokt kan betrokken worden in de afweging om reacties daarop als *hate speech* aan te brengen.[59]

In het algemeen kan worden gesteld dat het bij de inzet van zelflerende algoritmen moeilijker is om achteraf te bepalen hoe een specifieke beslissing of inschatting tot stand is gekomen dan bij regelgebaseerde algoritmen. Bovendien vermindert deze inzichtelijkheid bij meer geavanceerde *machine learning* technieken zoals *deep learning*. Voor modereeralgoritmen kan de inzichtelijkheid in de totstandkoming van de uitkomst afnemen naarmate meer randgegevens, zoals gegevens over de gebruiker, worden aangewend om uiteindelijk tot een inschatting te komen. In de grote

[54] Kaye 2018, p. 18.
[55] Gillespie 2018, p. 103.
[56] Schmidt & Wiegand 2017, p. 3.
[57] Schmidt & Wiegand 2017, p. 5.
[58] Gillespie 2018, p. 104. Zie ook Cheng, Danescu-Niculescu-Mizil & Leskovec 2015. Zie ook Mishra e.a. 2018, p. 1088: '*[p]revious research suggests that [...] abusive content tends to come from users who share a set of common stereotypes and form communities around them*.'
[59] Schmidt & Wiegand 2017, p. 6.

hoeveelheid data die dan wordt aangewend kunnen steeds moeilijker de doorslaggevende factoren worden aangewezen, aan de hand waarvan de beslissing zou kunnen worden verklaard.

Van zelfstandige besluitvorming door algoritmen ten aanzien van mogelijke *hate speech* op platformen is voor zover wij weten geen sprake. Bij platformen zoals Facebook en YouTube beslist uiteindelijk een medewerker van een platform over de door het algoritme aangebrachte content. De bescheiden rol van het algoritme in het besluitvormingsproces is te verklaren door de complexiteit van de afwegingen die moeten worden genomen en het belang van context voor de beoordeling van een uiting.[60] Wel is het mogelijk dat algoritmen worden ingezet om reeds als *hate speech* aangemerkte content te identificeren als die opnieuw gedeeld worden en de content dan automatisch te verwijderen.[61]

Hoewel er doorgaans uiteindelijk menselijke moderators beslissen over het verwijderen van content, staat of valt de juistheid van beslissingen bij de zorgvuldigheid die zij betrachten in het besluitvormingsproces. In dat verband is van belang dat de zuiverheid van de beoordeling kan lijden onder zowel de werkdruk waaronder contentmoderators opereren als de emotionele en psychische stress die dit werk met zich meebrengt.[62] Ook de subjectiviteit van de beoordelaar en diens opvattingen kunnen ertoe leiden dat bepaalde content niet altijd goed wordt beoordeeld.

4.2.3 Blik op de toekomst

In ons onderzoek zijn we geen platformen tegengekomen die het detecteren, beoordelen en verwijderen van *hate speech* in zijn geheel overlaten aan algoritmen.[63] Een toekomstbeeld dat zich in de nabije toekomst zou kunnen voltrekken is dat content, waarover bij het algoritme nauwelijks twijfel bestaat dat er sprake is van *hate speech*, automatisch wordt verwijderd of in quarantaine wordt gezet.[64] Er is dan sprake van zogenaamde ‘semi-automatische’ contentmoderatie. Platformen zouden het probleem van *hate speech* dan afkunnen met minder menselijke moderators, die alleen zouden hoeven te beslissen over de in quarantaine geplaatste twijfelgevallen.

Op de langere termijn is het denkbaar dat algoritmen een grotere rol gaan spelen in het *voorkomen* van *hate speech*. Zo kunnen algoritmen gaan reageren op uitingen van *hate speech* in een poging

[60] Zie interview van TechCrunch met Timothy Quinn, CEO van Hatebase (Coldewey, techcrunch.nl 10 september 2019).
[61] Zie in dat verband ook HvJ EU 3 oktober 2019, ECLI:EU:C:2019:821 (*Eva Glawischnig-Piesczek/Facebook*).
[62] Zie in dat verband bijvoorbeeld Roberts 2016; Roberts 2014.
[63] Instagram is een voorbeeld van een platform dat met betrekking tot reacties die worden geplaatst op posts wel automatisch filtert met behulp van algoritmen. Instagram heeft een algoritme getraind dat ‘*offensive comments*’ op Instagramposts kan herkennen en verbergen. Strikt genomen gaat het hier niet om contentmoderatie door platformen, maar om een functionaliteit die door gebruikers van Instagram vrijwillig aangezet kan worden met betrekking tot de reacties die zij ontvangen op hun posts. Zie daarover Instagram, ‘Keeping Instagram a Safe Place for Self-Expression’, 29 juni 2017, instagram-press.com/blog/2017/06/29/keeping-instagram-a-safe-place-for-self-expression/. Zie ook Roberts, *VICE* 2 mei 2018.
[64] Pavlopoulos, Malakasiotis & Androutsopoulos 2017.

de ‘uiter’ te bewegen zijn gedrag in de toekomst aan te passen.[65] Mensen kunnen er, met behulp van algoritmen, in dit verband mogelijk ook toe worden bewogen om een concrete uiting niet te posten of die aan te passen.[66] Een gevaar is dan dat mensen onvoldoende vrij zijn zichzelf te uiten en dat de inhoud en de grenzen van het publieke debat mede worden bepaald door algoritmen.

4.3 Vrijheid van meningsuiting en informatie

Het verbieden van een bepaalde categorie uitingen en het handhaven daarvan staan in direct verband met de publieke waarde van vrijheid van meningsuiting, die op zichzelf weer een belangrijke pijler vormt onder onze democratische rechtsstaat. Hieronder wordt eerst de vrijheid van meningsuiting als publieke waarde geïntroduceerd en de wijze waarop die beschermd wordt. Daarna gaan we met betrekking tot het modereren van *hate speech* nader in op de kansen en risico’s voor de vrijheid van meningsuiting en evalueren we het relevante juridisch kader.

4.3.1 Vrijheid van meningsuiting en contentmoderatie

De vrijheid van meningsuiting omvat niet alleen het hebben en delen van een bepaalde mening, maar ook het ontvangen van meningen en informatie van anderen. Het modereren van online *hate speech* raakt daarmee aan zowel de vrijheid van gebruikers om hun mening te uiten en informatie te delen, als de vrijheid van ontvangers om meningen en informatie tot zich te nemen. In Nederland is de bescherming van de vrijheid van meningsuiting verankerd in art. 7 Gw, art. 10 EVRM en art. 11 Handvest. Ook choquerende en beledigende uitingen worden beschermd door de vrijheid van meningsuiting, omdat de belangrijke waarden in de democratische samenleving van pluralisme, tolerantie en ruimdenkendheid dat eisen.[67]

De vrijheid van meningsuiting is echter niet absoluut. Hoewel het belangrijk is dat schokkende of politiek gevoelige meningen bescherming genieten, betekent dit geen *carte blanche* voor haatdragende of haatzaaiende uitingen.[68] De wetgever kan de vrijheid van meningsuiting dan ook inperken ten behoeve van de bescherming van bepaalde publieke belangen en de bescherming van andere grondrechten en rechten van anderen. Een dergelijke inperking van de vrijheid van meningsuiting moet bij wet zijn voorzien, noodzakelijk zijn in een democratische samenleving, en

[65] Zie voor een voorbeeld waarin een Twitterbot racistische uitingen ‘sanctioneert’: Munger, *Political Behavior* 2017, p. 629-649.
[66] Zie voor een voorbeeld Santos e.a. 2018. Zie Jurgens, Chandrasekharan & Hemphill 2019 voor een oproep voor een proactieve inzet van *natural language processing* technologieën.
[67] EHRM 7 december 1976, ECLI:CE:ECHR:1976:1207JUD000549372 (*Handyside/Verenigd Koninkrijk*), r.o. 49.
[68] In dat verband moet ook worden opgemerkt dat het EHRM in bepaalde gevallen van *hate speech* een beroep op art. 10 EVRM niet-ontvankelijk of kennelijk ongegrond heeft verklaard, omdat de uitingen in kwestie de rechten van het EVRM aanvallen of ondermijnen. Zie in dat verband Keane, *Netherlands Quarterly of Human Rights* 2007, p. 641.

moet proportioneel zijn aan het nagestreefde legitieme doel. Zo kunnen ook bepaalde vormen van *hate speech* aan banden worden gelegd.[69]

Platformen zijn vrijgesteld van civielrechtelijke aansprakelijkheid voor opgeslagen content zolang zij niet weten dat er onrechtmatige content op hun servers staat.[70] In het geval zij kennis verkrijgen van onmiskenbaar onrechtmatige content, dienen platformen die onrechtmatige content 'prompt' te verwijderen.[71] Platformen hebben er daarom belang bij om onrechtmatige content te verwijderen als zij daarvan op de hoogte worden gesteld, omdat zij zich anders niet op de aansprakelijkheidsbeperking kunnen beroepen. Met deze regels wordt een balans gezocht tussen enerzijds de rechtszekerheid van internetdienstverleners met betrekking tot hun aansprakelijkheidspositie en anderzijds het belang om de verspreiding van illegale informatie en onrechtmatige activiteiten op het internet tegen te gaan.[72]

Uit art. 6:196c lid 5 BW volgt dat platformen, ondanks de aansprakelijkheidsvrijwaring, een verbod of bevel opgelegd kunnen krijgen om de verspreiding van onrechtmatige content te stoppen of te voorkomen. Wanneer een dergelijk verbod of bevel uitgevaardigd moet worden, wordt niet door het Europees recht bepaald. Art. 15 van de E-Commercerichtlijn stelt echter een grens aan wat er verlangd mag worden van platformen.[73] Verplichtingen mogen niet strekken tot een 'algemene toezichtverplichting' waarbij platformen alle geüploade materialen controleren op mogelijke illegale informatie of onrechtmatige activiteiten. Uit jurisprudentie van het HvJ EU volgt echter dat art. 15 van de E-Commercerichtlijn niet in de weg staat aan een rechterlijk bevel om reeds als onrechtmatige aangemerkte specifieke content ook in de toekomst verwijderd te houden.[74] Het is naar EU-recht ook geoorloofd om een bevel uit te vaardigen dat ertoe strekt ook 'overeenstemmende' content verwijderd te houden, voor zover het platform 'geautomatiseerde technieken en onderzoeksmethoden' kan toepassen en de content geen autonome beoordeling behoeft.[75] Het is daardoor denkbaar dat een Nederlandse rechter in de toekomst beslist dat een platform dient te voorkomen dat een specifieke of overeenstemmende onrechtmatige uiting op het

[69] Zie ook EHRM 6 juli 2006, ECLI:CE:ECHR:2006:0706JUD005940500 (*Erbakan/Turkije*), r.o. 56. Ook de aansprakelijkheid van bepaalde online dienstverleners voor dergelijke uitingen kan volgens het EHRM de toets van art. 10 EVRM doorstaan, zie EHRM 16 juni 2015, ECLI:CE:ECHR:2015:0616JUD006456909 (*Delfi AS/Estland*); EHRM 2 februari 2016, ECLI:CE:ECHR:2016:0202JUD002294713 (*MTE & Index.hu ZRT/Hongarije*). De uitspraken laten zich echter niet goed verhouden tot het Unierechtelijke kader waar dergelijke aansprakelijkheid is uitgesloten.
[70] Deze aansprakelijkheidsbeperkingen gelden ook voor eventuele strafrechtelijke aansprakelijkheid.
[71] *Kamerstukken II* 2003/04, 28197, nr. 15, p. 2.
[72] Nederland kent sinds 2008 een Gedragscode Notice-and-Takedown die wordt onderhouden door het Platform voor de InformatieSamenleving. Bij deze gedragscode zijn traditionele hostingbedrijven, maar ook platformen zoals Google en Facebook aangesloten. De gedragscode biedt een procedure voor notice-and-takedown en probeert onzekerheden weg te nemen ten aanzien van de te volgen procedure, de voorwaarden waaronder verwijderd dient te worden, en de timing van verwijdering. De inzet van algoritmen bij het detecteren van content door middel van algoritmen wordt echter niet gereguleerd door deze code. Zie voor de code: noticeandtakedowncode.nl.
[73] Richtlijn 2000/31/EG van het Europees Parlement en de Raad van 8 juni 2000 betreffende bepaalde juridische aspecten van de diensten van de informatiemaatschappij, met name de elektronische handel, in de interne markt (*PbEG* 2000, L 178/1. Art. 15 van deze richtlijn is niet expliciet geïmplementeerd in het Nederlandse recht.
[74] HvJ EU 3 oktober 2019, ECLI:EU:C:2019:821 (*Eva Glawischnig-Piesczek/Facebook*).
[75] HvJ EU 3 oktober 2019, ECLI:EU:C:2019:821 (*Eva Glawischnig-Piesczek/Facebook*).

platform wordt geplaatst. Een dergelijke beslissing zou ertoe kunnen leiden dat een platform algoritmen moet inzetten om die content te (helpen) detecteren en te verwijderen.

Specifiek ten aanzien van de aanpak van illegale *hate speech* door online platformen is er op Europees niveau de in 2016 opgestelde *Code of Conduct on Countering Illegal Hate Speech Online*. In die code heeft de Europese Commissie afspraken gemaakt met een aantal grote online platformen over de bestrijding van illegale online *hate speech*.[76] Daarin wordt onder andere gestreefd naar snelle en effectieve verwijdering van illegale *hate speech* naar aanleiding van meldingen.[77] Met de code verbinden de betrokken platformen zich ertoe de vrijheid van meningsuiting te bevorderen en faciliteren.[78] De code onderstreept ook het belang van de bestrijding van *hate speech*: '*The spread of illegal hate speech online not only negatively affects the groups or individuals that it targets, it also negatively impacts those who speak out for freedom, tolerance and non-discrimination in our open societies and has a chilling effect on the democratic discourse on online platforms.*'[79]

Naast deze specifieke code ten aanzien van *hate speech*, is ook de aanbeveling van de Europese Commissie met betrekking tot '*Measures to Effectively Tackle Illegal Content Online*' relevant.[80] Deze aanbeveling roept hosting providers, waaronder online platformen, op om effectieve, geschikte en proportionele maatregelen te nemen om de verspreiding van *illegal content* tegen te gaan. De aanbeveling roept providers op om ook proactief op te treden tegen illegale content, daar waar dat gepast en proportioneel is, onder meer door algoritmen in te zetten voor het detecteren van illegale content.[81] In de aanbeveling wordt bovendien benadrukt dat in alle gevallen waarin content verwijderd wordt of de toegang ertoe geblokkeerd, online platformen zich bewust moeten tonen van de centrale rol die zij hebben voor het faciliteren van het publieke debat, en voor het verspreiden en ontvangen van feiten, meningen en ideeën.[82]

[76] Microsoft, Twitter, Youtube, Instagram, (het inmiddels opgeheven) Google+, Snapchat, Dailymotion en Jeuxvideo.com zijn aangesloten bij de code. Zie voor de code en rapporten met betrekking tot de monitoring van de code 'The EU Code of conduct on countering illegal hate speech online. The robust response provided by the European Union', ec.europa.eu.
[77] Code of Conduct on Countering Illegal Hate Speech Online, 30 juni 2016, p. 1: '*The IT Companies to review the majority of valid notifications for removal of illegal hate speech in less than 24 hours and remove or disable access to such content, if necessary*.'
[78] Code of Conduct on Countering Illegal Hate Speech Online, 30 juni 2016, p. 1: 'Facebook*, Microsoft*, Twitter and YouTube (hereinafter "the IT Companies") – also involved in the EU Internet Forum – share, together with other platforms and social media companies, a collective responsibility and pride in promoting and facilitating freedom of expression throughout the online world.*'
[79] Code of Conduct on Countering Illegal Hate Speech Online, 30 juni 2016, p. 1.
[80] C(2018) 1177 final. In de aanbeveling wordt *illegal content* gedefinieerd als '*any information which is not in compliance with Union law or the law of a Member State concerned*'. Hoewel de aanbevelingen zich lijken toe te spitsen op informatie die naar haar aard onrechtmatig is, wordt er in de overwegingen bijvoorbeeld ook gewezen op de bestrijding van ongeautoriseerde verspreiding van auteursrechtelijk beschermd werk.
[81] C(2018) 1177 final, punt 18. De aanbeveling spreekt hier van '*use of automated means for the detection of illegal content*'.
[82] C(2018) 1177 final, overweging 31.

4.3.2 Kansen, risico's en bestendigheid juridisch kader

De kansen die de inzet van contentmodereeralgoritmen biedt voor de vrijheid van meningsuiting hebben met name betrekking op de bijdrage die zij kunnen leveren aan een pluralistische, tolerante en open samenleving. Contentmodereeralgoritmen vergroten de mogelijkheden om van online platformen een veiligere plek te maken voor met name minderheden om deel te nemen aan het publieke debat. Zij kunnen daarmee ook helpen voorkomen dat over bepaalde onderwerpen het debat niet wordt aangegaan omdat men bang is slachtoffer te worden van *hate speech*.[83]

Als het gaat om de bestrijding van *hate speech* dan is een grote uitdaging hoe en volgens welke maatstaven in concrete gevallen wordt vastgesteld of sprake is van *hate speech* en hoe daarbij kan worden voorkomen dat rechtmatige uitingen worden gecensureerd.[84] De schaal waarop in het online domein content moet worden gemodereerd, vergroot die uitdaging alleen maar. Algoritmen kunnen in dat opzicht een belangrijke bijdrage leveren, omdat ze het mogelijk maken om ook op grote schaal content te modereren.

Risico's die voortvloeien uit de inzet van algoritmen ten aanzien van de vrijheid van meningsuiting bestaan hoofdzakelijk in het onterecht aanbrengen van uitingen als *hate speech*. De oorzaken daarvan zijn reeds aangestipt in par. 4.2.2. Hoewel de uiteindelijke beslissing ten aanzien van de toelaatbaarheid van een uiting aan een menselijke content-moderator is, kunnen vooroordelen en andere omstandigheden die leiden tot onterechte kwalificatie als mogelijke *hate speech* door het algoritme toch een rol spelen als de menselijke moderator onder hoge druk snelle beslissingen moet nemen. Omgekeerd kunnen deze omstandigheden een rol spelen bij het onterecht niet kwalificeren van content als *hate speech*, wat de bestrijding van *hate speech* weer afhankelijk maakt van gebruikers die de content moeten rapporteren.

De *Code of Conduct on Countering Illegal Hate Speech Online*[85] is een door de EU geïnitieerd zelfreguleringsinstrument dat voorziet in concrete stappen die platformen moeten nemen om *hate speech* terug te dringen. Daarin wordt gestreefd naar snelle en effectieve verwijdering naar aanleiding van meldingen. Ook spreken platformen uit dat zij ernaar streven om onder andere inzicht te geven in de procedure die wordt gevolgd voor behandeling van meldingen, training te geven aan hun medewerkers, en voorlichting te geven aan gebruikers met betrekking tot wat er is toegestaan op het platform.

[83] Zie voor een Deense studie naar het verband tussen online speech en de deelname aan het debat, ook in relatie tot bepaalde onderwerpen Zuleta & Burkal 2017. Ook de Duitse *Netzwerkdurchsetzungsgesetz* (NetzDG) tracht zo een positieve bijdrage te leveren aan de vrijheid van meningsuiting, zie Theil, *Verfassungsblog* 8 februari 2018. Zie ook in algemene zin met betrekking tot de bestrijding van *hate speech* McGonagle 2013, p. 6.
[84] Massaro, *William and Mary Law Review* 1991, p. 214-215.
[85] Zie voetnoot 39 hierboven.

De implementatie van dit instrument wordt gemonitord door de Europese Commissie. De nadruk in de monitoring ligt daarin steeds op de '*removal rates*' van *hate speech*. Naleving en de effectiviteit van de code wordt met name uitgedrukt in en gemeten aan de hoeveelheden en percentages van *hate speech* die worden verwijderd naar aanleiding van gedane meldingen, en de snelheid waarmee dit gebeurt. De correctheid en de zorgvuldigheid van de besluitvorming komt echter nauwelijks ter sprake in de rapportages over de monitoring. Daarnaast valt op dat, hoewel de code aandacht besteedt aan de rol die onderwijs kan spelen bij het terugdringen van *hate speech*, er in de monitoring daarvoor ook geen aandacht is. Omdat de code en de monitoring ervan slechts betrekking hebben op het verwerken van meldingen van *hate speech* en niet op de inzet van algoritmen om zelf *hate speech* te detecteren, kan de code de risico's die samenhangen met de inzet van algoritmen voor de vrijheid van meningsuiting niet ondervangen.

In de algemenere Aanbeveling van de Europese Commissie met betrekking tot '*Measures to Effectively Tackle Illegal Content Online*' is er wel aandacht voor het proactief modereren van content met behulp van algoritmen. Een vraag die rijst is in hoeverre een dergelijk niet-bindend instrument platformen ertoe beweegt om het modereren van content vergezeld te doen gaan van waarborgen waarmee de vrijheid van meningsuiting voldoende kan worden beschermd.[86]

In de Aanbeveling is desalniettemin aandacht voor het gevaar dat rechtmatige content wordt verwijderd. Ook de rol die algoritmen daarin kunnen spelen komt aan de orde. Als hosting providers gebruikmaken van *automated means* om content te analyseren, dan dienen gepaste waarborgen te worden geboden om ervoor te zorgen dat genomen beslissingen precies (*'accurate'*) en op terechte ('*well-founded*') gronden worden genomen, in het bijzonder als het gaat om besluiten om content te verwijderen of de toegang ertoe te blokkeren.[87] Deze waarborgen zouden, waar gepast, moeten bestaan in het uitoefenen van toezicht ('*human oversight*') en het verrichten van verificaties door mensen. Dat zou in ieder geval moeten gebeuren waar een gedetailleerde beoordeling van de relevante context nodig is om te bepalen of bepaalde content illegaal is. Ook het menselijke toezicht dat gehouden moet worden, is niet nader aan voorwaarden verbonden.

De Aanbeveling roept ook op tot het geven van inzage in het beleid en de praktijken met betrekking tot het verwijderen van illegale content. Hosting providers worden aangemoedigd om met enige regelmaat daarover rapporten te publiceren, in het bijzonder ten aanzien van de hoeveelheid en het type content dat is verwijderd, de aantallen meldingen en bezwaren die zijn ingediend en de

[86] In dat verband is het tekenend dat de nieuwe Europese Commissie onder leiding van Von der Leyen heeft aangekondigd de aansprakelijkheid van online dienstverleners nader te willen regelen in wetgeving. Zie daarover Von der Leyen 2019, p. 13.
[87] C(2018) 1177 final, punt 19.

tijd die het heeft gekost om actie te ondernemen. De code roept niet specifiek op tot inzage in de wijze waarop algoritmen een rol spelen in het modereren van content.[88]

4.3.3 Tussenconclusie

Het reguleren en handhaven van online *hate speech* is onlosmakelijk verbonden met de publieke waarde van de vrijheid van meningsuiting. Hoewel terughoudendheid belangrijk is bij het verbieden en verwijderen van bepaalde uitingen, kan dit in bepaalde gevallen de vrijheid van meningsuiting ook ten goede komen omdat daarmee de voorwaarden worden geschept voor een inclusieve online omgeving die uitnodigt tot participatie. De inzet van algoritmen voor het detecteren van *hate speech* kan daaraan bijdragen, maar brengt ook risico's met zich mee. Als algoritmen uitingen onterecht als mogelijke *hate speech* aanmerken, en het menselijke toezicht onvoldoende is om dat recht te zetten, dan kan de inzet van algoritmen leiden tot censuur. In de bestaande juridische kaders is daarvoor echter onvoldoende aandacht.

De grondrechten, zoals de vrijheid van meningsuiting, kunnen handvatten bieden om deze problemen aan te pakken. Een complicatie in dat verband is wel dat het modereren van *hate speech* plaatsheeft op private platformen die zich door middel van *Terms of Service* en huisregels in een privaatrechtelijke relatie tot hun gebruikers verhouden. De grondrechten zijn niet direct van toepassing in deze verhoudingen. Wel kan gesteld worden dat overheden een positieve verplichting hebben om de vrijheid van meningsuiting te waarborgen als platformen content modereren.[89] Daaruit zou dan weer kunnen volgen dat overheden bestaande kaders dienen aan te passen om de risico's van algoritmische besluitvorming ten aanzien van de vrijheid van meningsuiting op effectieve wijze te mitigeren.

4.4 Bescherming van persoonsgegevens

4.4.1 Bescherming van persoonsgegevens en contentmoderatie

De bescherming van persoonsgegevens kan in het geding komen als algoritmen, naast een inhoudelijke analyse van de content, ook informatie over de gebruiker betrekken in hun beoordeling van de content. Een van de uitgangspunten van het gegevensbeschermingsrecht is dat individuen controle houden over de verwerking van hun persoonsgegevens.

4.4.2 Kansen, risico's en bestendigheid juridisch kader

Risico's ten aanzien van de bescherming van persoonsgegevens kunnen met name ontstaan als online platformen profielen bouwen van gebruikers, bijvoorbeeld met betrekking tot het type content dat zij posten en de kans dat die content onrechtmatig of onwenselijk is. Dergelijke

[88] Dat is wel het geval als het gaat om terroristische content, zie C(2018) 1177 final, punt 42.
[89] Zie in dat verband Angelopoulos e.a. 2015.

classificaties kunnen leiden tot besluitvorming waarbij stereotypes centraal komen te staan, en brengen, afhankelijk van het soort gegevens dat wordt gebruikt, mogelijk een verhoogd gevaar van discriminatie en onjuiste besluitvorming mee (zie ook par. 4.5). Ook worden er dan mogelijk gevoelige gegevens verzameld, die in geval van misbruik kunnen leiden tot schade aan de reputatie van gebruikers.[90]

De Algemene Verordening Gegevensbescherming (AVG), maar ook de gelijkheidswetgeving, biedt diverse waarborgen tegen voornoemde risico's. In de AVG zijn diverse beginselen neergelegd, zoals de beginselen van dataminimalisatie, doelbinding en transparantie, die ook zijn uitgewerkt in nadere regels in de AVG.[91] Als online content wordt gemodereerd en daarbij persoonsgegevens worden betrokken, dan dient aan die regels en beginselen te worden voldaan.[92]

4.4.3 Tussenconclusie

De bescherming van persoonsgegevens kan mogelijk in het geding komen als platformen gegevens over gebruikers betrekken in de analyse van content. Een risico dat zich kan voordoen is het overmatig gebruik van persoonsgegevens in een poging om het modereren van online content te optimaliseren. De beginselen en regels in de AVG bieden voldoende aanknopingspunten om gebruik van persoonsgegevens door online platformen te reguleren.

4.5 Non-discriminatie

4.5.1 Non-discriminatie en contentmoderatie

De aanpak van *hate speech* is nauw verbonden met de publieke waarde van non-discriminatie, in zoverre dat *hate speech* veelal de (gelijk)waardigheid van groepen personen ondermijnt of er zelfs op is gericht die te ondermijnen.[93] Racisme, seksisme en andere houdingen die discriminatie tussen groepen personen in stand houden of in de hand werken, kunnen door *hate speech* worden bestendigd.[94] De eerder genoemde *Code of Conduct on Countering Illegal Hate Speech Online*[95] beoogt daarom online *hate speech* op platformen te bestrijden.

[90] Zie met betrekking tot de risico's van profilering in algemene zin Schermer 2013, p. 137.
[91] Zie voor de beginselen art. 5 AVG.
[92] Zie in dat verband ook C(2018) 1177 final, overwegingen 13 en 39, waarin wordt benadrukt dat bij de aanpak van illegale online content onder andere het grondrecht op gegevensbescherming moet worden gerespecteerd, en dat maatregelen die worden genomen om gevolg te geven aan de Aanbeveling volledig in overeenstemming moeten zijn met de geldende regels ten aanzien van gegevensbescherming.
[93] Waldron 2012.
[94] Cowan, *Journal of Social Issues* 2002, p. 250.
[95] Zie voetnoot 39 hierboven.

4.5.2 Kansen, risico's en bestendigheid juridisch kader

De inzet van algoritmen voor het detecteren van *hate speech* biedt kansen voor de aanpak van discriminatie in de zin dat racistische, seksistische en andere haatzaaiende of haatdragende uitingen sneller kunnen worden gedetecteerd en daartegen kan worden opgetreden, wat zou moeten leiden tot inclusievere online omgevingen.

Tegenover die kans staat het aanmerkelijke risico dat de algoritmen die hiervoor worden ingezet de vooringenomenheden van hun programmeurs overnemen, of vooroordelen weerspiegelen die aanwezig zijn in de data waarmee deze algoritmen worden getraind.[96] Twee recente studies wijzen er bijvoorbeeld op dat de modereeralgoritmen van Twitter content vaker als beledigend of als haatzaaiend aanmerken wanneer deze elementen van straattaal of '*slang*' bevat.[97] Deze vooringenomenheid van het algoritme is te herleiden tot vooroordelen die leven ten aanzien van mensen die straattaal bezigen. Als ook menselijke moderators ten aanzien van content met straattaal besluiten dat er sneller sprake is van *hate speech,* en die besluiten weer de invoer vormen voor het aanscherpen van het algoritme, dan kan er een *feedback loop* ontstaan waarbij de ongelijke behandeling dieper wordt verankerd. De risico's die bestaan ten aanzien van de vrijheid van meningsuiting gelden dan ook sterker voor groepen personen waarover al vooroordelen bestaan. De impact van een dergelijk discriminerend effect kan bovendien worden versterkt als gevolg van de schaalvergroting die mogelijk wordt door de inzet van algoritmen.

Tegelijkertijd is het denkbaar dat juist met algoritmen de vooringenomenheden en vooroordelen in modereeralgoritmen kunnen worden blootgelegd. Waar de inzet van modereeralgoritmen leidt tot vormen van discriminatie van bepaalde beschermde groepen, zijn zulke algoritmen bijzonder waardevol. Het verbod op verwerking van bijzondere persoonsgegevens in de AVG staat echter mogelijk in de weg aan de succesvolle implementatie van zulke algoritmen, aangezien daarvoor juist het gebruik van bijzondere persoonsgegevens zoals ethniciteit en ras nodig kan zijn.[98]

Zowel de *Code of Conduct* als de Aanbevelingen van de Europese Commissie bevatten geen specifieke normen die strekken tot het voorkomen van vooroordelen in modereeralgoritmen en de bovengenoemde mogelijkheden van discriminatie. Daarnaast worden platformen niet opgeroepen om zelf *best practices* te ontwikkelen waarmee de risico's op discriminatie zouden kunnen worden ondervangen. Ook aan het menselijke toezicht dat gehouden moet worden, worden geen nadere voorwaarden gesteld die vooringenomenheid of discriminatie kunnen tegengaan.

[96] Binns e.a. 2017.
[97] Davidson, Bhattacharya & Weber 2019; Sap e.a. 2019. Zie ook, met betrekking tot een dataset van Wikipedia, Binns e.a. 2017.
[98] Art. 9 AVG.

4.5.3 Tussenconclusie

De inzet van algoritmen kan mogelijk bijdragen aan een inclusievere online omgeving waarin discriminatie op een efficiëntere wijze wordt bestreden. Echter, als vooringenomenheden de ingezette modereeralgoritmen binnensluipen, levert dat ook weer gevaren op voor de publieke waarde van non-discriminatie. De *Code of Conduct on Countering Illegal Hate Speech Online*[99] en de Aanbeveling van de Europese Commissie met betrekking tot '*Measures to Effectively Tackle Illegal Content Online*'[100] laten deze problematiek onbesproken, waardoor deze instrumenten geen of nauwelijks richting bieden ten aanzien van het voorkomen of anderszins mitigeren van mogelijke discriminatie door modereeralgoritmen.

4.6 Rechtsbescherming

4.6.1 Rechtsbescherming en contentmoderatie

Personen die door hen geplaatste content verwijderd zien, moeten de beslissing die daaraan ten grondslag ligt, kunnen aanvechten. Daarvoor is ook van belang dat zij begrijpen op grond waarvan hun content wordt verwijderd.

4.6.2 Kansen, risico's en bestendigheid juridisch kader

Een lastigheid is dat de relatie tussen online platformen en hun gebruikers en de beschikbaarheid van eventuele bezwaarmogelijkheden worden beheerst door overeenkomsten zoals *Terms of Service*. Omdat platformen zoals YouTube en Facebook als dominant zijn aan te merken, is van een gelijkwaardige verhouding tussen platform en gebruikers, en een vrije beslissing over het al dan niet accepteren van de voorwaarden, feitelijk echter geen sprake. Dit resulteert onder andere in een algemeen gebrek aan (serieuze) mogelijkheden voor gebruikers om bezwaar te maken tegen de beslissingen van platformen. De negatieve impact van dat gebrek op de rechtsbescherming van gebruikers geldt mogelijk nog sterker wanneer voor zulke besluitvorming ook algoritmen worden ingezet. Omdat modereeralgoritmen door allerlei factoren tot uitkomsten kunnen komen die mensen onlogisch of duidelijk verkeerd voorkomen, wordt het probleem van het gebrek aan bezwaarmogelijkheden nog prangender.[101]

Een risico met betrekking tot de rechtsbescherming van gebruikers van platformen is de gebrekkige inzichtelijkheid van het besluitvormingsproces en de rol die algoritmen daarin spelen op zowel individueel als meer algemeen niveau. Voor rechtsbescherming van platformgebruikers is het van belang dat zij begrijpen op grond waarvan hun uitingen zijn verwijderd en dat hen

[99] Zie voetnoot 39 hierboven.
[100] C(2018) 1177 final.
[101] Zie ook Meyers West, *New Media & Society* 2018, p. 4366, over het onbegrip dat bestaat bij gebruikers over besluiten die online platformen maken op het gebied van contentmoderatie.

daarover informatie wordt verstrekt. Dat vergt ook algoritmen die tot een uitlegbare uitvoer kunnen komen.[102] Als zelflerende algoritmen worden ingezet dan is dat echter geen eenvoudige opgave.[103]

Inzichtelijkheid in de werking van modereeralgoritmen in algemene zin kan eraan bijdragen dat duidelijk wordt hoe groot de risico's zijn met betrekking tot de vrijheid van meningsuiting en non-discriminatie. Als dat inzicht wordt verkregen dan wordt het mogelijk om daarop beleid te voeren en regulering daarop toe te snijden.

In dat kader wreekt zich dat de *Code of Conduct on Countering Illegal Hate Speech Online*[104] is toegespitst op het afhandelen van meldingen over content die is aangebracht door andere gebruikers en niet zozeer door modereeralgoritmen. De Aanbeveling '*Measures to Effectively Tackle Illegal Content Online*'[105] bevat wel bepalingen die raken aan de rechtsbescherming van de gebruiker. Hij dient op de hoogte te worden gebracht van de redenen om zijn content te verwijderen.[106] De gebruiker moet ook een mogelijkheid hebben om bezwaar aan te tekenen.[107] Platformen dienen volgens de Aanbeveling ook in algemene zin inzage te geven in het beleid en de praktijken met betrekking tot het verwijderen van illegale content. Platformen worden aangemoedigd om met enige regelmaat daarover rapporten te publiceren, in het bijzonder ook ten aanzien van de hoeveelheid en het type content dat is verwijderd, de aantallen meldingen en bezwaren die zijn ingediend en de tijd die het heeft gekost om actie te ondernemen.[108] De Code roept eveneens niet specifiek op tot inzage in de wijze waarop algoritmen een rol spelen in het modereren van content.[109]

Art. 22 AVG kan ook een rol spelen in de rechtsbescherming van platformgebruikers. Dit artikel bepaalt dat personen wier persoonsgegevens worden verwerkt het recht hebben om 'niet te worden onderworpen aan een uitsluitend op geautomatiseerde verwerking, waaronder profilering, gebaseerd besluit waaraan voor hem rechtsgevolgen zijn verbonden of dat hem anderszins in aanmerkelijke mate treft'.

Art. 22 AVG lijkt echter niet van toepassing op de huidige modereerpraktijk van online platformen. Het is in de eerste plaats onzeker of er in geval van het verwijderen van content sprake is van een beslissing waaraan rechtsgevolgen zijn verbonden, dan wel een beslissing die de gebruiker van een platform in aanmerkelijke mate treft. Afhankelijk van de uiting in kwestie en de gevolgen die niet-plaatsen van de content heeft, zou er mogelijk sprake kunnen zijn van een beslissing die de

[102] Zie ook Suzor e.a., *International Journal of Communication* 2019, p. 1526.
[103] Zie par. 2.1.2.
[104] Zie voetnoot 39 hierboven.
[105] C(2018) 1177 final.
[106] C(2018) 1177 final, punt 9.
[107] C(2018) 1177 final, punt 11.
[108] C(2018) 1177 final, punt 17.
[109] Dat is wel het geval als het gaat om terroristische content, zie C(2018) 1177 final, punt 42.

gebruiker in aanmerkelijke mate treft omdat daardoor zijn vrijheid van meningsuiting wordt beperkt. Echter, voor zover er sprake is van een dergelijke beslissing, dan geldt dat de beslissing waarschijnlijk geen besluit is dat is gebaseerd op een *uitsluitend* geautomatiseerde verwerking. Het modereren van *hate speech* vergt op dit moment namelijk nog steeds inmenging van een menselijke moderator die uiteindelijk beslist over en verantwoordelijkheid heeft voor het wel of niet verwijderen van content.[110]

Het kan niet worden uitgesloten dat het modereren van *hate speech* in de toekomst verder wordt geautomatiseerd en menselijke tussenkomst niet langer nodig is. Art. 22 AVG is dan, zeker als de besluitvorming discriminerende effecten heeft, waarschijnlijk wel van toepassing.[111] Platformen zouden dan de uitdrukkelijke toestemming moeten verkrijgen van platformgebruikers voor de inzet van modereeralgoritmen.[112] Voor platformen is dan van belang dat zij maatregelen nemen ter bescherming van de rechten en vrijheden van hun gebruikers. Daartoe behoren ook het recht op menselijke tussenkomst, het recht van de gebruiker om zijn standpunt kenbaar te maken en het recht om het modereerbesluit aan te vechten.[113] Ook het recht op informatie met betrekking tot de geautomatiseerde besluitvorming is dan van toepassing.[114] Dat zou de gebruiker van een online platform ook het recht geven om 'nuttige informatie over de onderliggende logica' van het besluitvormingsproces te ontvangen. Dat zou van online platforms vergen dat zij begrijpelijke uitleg geven over de waarschijnlijk zeer complexe wijze waarop een algoritme beslist.[115]

4.6.3 Tussenconclusie

Rechtsbescherming ondersteunt de andere relevante publieke waarden en kan eraan bijdragen dat eventuele risico's van de inzet van algoritmen door platforms met betrekking tot de vrijheid van meningsuiting en non-discriminatie kunnen worden gemitigeerd of in ieder geval gepaard gaan met bepaalde waarborgen voor gebruikers. Inzicht in de rol die algoritmen spelen in het besluitvormingsproces kunnen daarbij van belang zijn. Het huidige juridisch kader biedt echter geen transparantieverplichtingen. Ook de Aanbeveling '*Measures to Effectively Tackle Illegal Content Online*' roept niet op tot het inzichtelijk maken van de werking van algoritmen die onrechtmatige content aanbrengen. Hoewel art. 22 AVG verschillende van zulke waarborgen biedt ten aanzien van de rechtsbescherming bij geautomatiseerde besluitvorming, moet worden geconstateerd dat de huidige contentmoderatiepraktijk in ieder geval op dit moment nog niet zodanig geautomatiseerd is dat die bepaling van toepassing is. Tot dat moment lijkt de rechtsbescherming van gebruikers bij contentmoderatie dan ook onvoldoende gewaarborgd.

[110] Vgl. Groep gegevensbescherming artikel 29 2017b, p. 24.
[111] Vgl. Groep gegevensbescherming artikel 29 2017b, p. 26.
[112] Art. 22 lid 2 onder c AVG.
[113] Art. 22 lid 3 AVG.
[114] Art. 13 lid 2 onder f en art. 14 lid 2 onder g AVG.
[115] Vgl. Groep gegevensbescherming artikel 29 2017b, p. 25.

4.7 Conclusie

De inzet van algoritmen die onrechtmatige of onwenselijke content kunnen opsporen en aanbrengen, is onlosmakelijk verbonden met de schaal waarop content vandaag de dag wordt gegenereerd en gedeeld.[116] 'Traditionele' methoden van modereren zijn niet toegerust op deze hoeveelheden content, zodat de automatisering van een deel van het modereerproces in het huidige internetlandschap onontkoombaar is. Een kans die daarom in algemene zin geldt voor alle vormen van contentmoderatie is dat online platformen met het gebruik van algoritmen het probleem van schaal waarop illegale en onwenselijke content wordt gedeeld het hoofd kunnen bieden.

Om de rol die algoritmen (kunnen) spelen in contentmoderatie en de daarmee samenhangende kansen en risico's te kunnen waarderen, dient onderscheid te worden gemaakt naar het type content in kwestie, de onrechtmatigheid of onwenselijkheid van die content, en de gevolgen ervan. De inzet van algoritmen voor het modereren van content zal afhankelijk van deze onderscheidingen meer of minder geschikt zijn. Zo zijn uitingen die naar hun aard onrechtmatig zijn waarschijnlijk eenvoudiger te detecteren dan uitingen waarvoor een uitgebreide analyse nodig is van de context waarin een bepaalde uiting is gedaan. Daarnaast kan ook content die reeds als onrechtmatig of onwenselijk is aangemerkt eenvoudiger gedetecteerd worden dan content waarvan dat (nog) niet vaststaat.

In deze casestudy bestudeerden wij in het bijzonder de inzet van algoritmen door online platformen voor het detecteren van *hate speech*. Algoritmen maken het mogelijk om op grote schaal *hate speech* aan te pakken en dat biedt kansen voor het inclusiever maken van online omgevingen, waarin personen zich niet ontmoedigd voelen om te participeren in het publiek debat. Online platformen nemen daarin al de nodige stappen. Ook de aansprakelijkheidsnormen en zelfreguleringsinstrumenten bieden stimulansen om *hate speech* proactief op te sporen met behulp van algoritmen.

Risico's die samenhangen met de inzet van algoritmen om *hate speech* te detecteren doen zich vooral voor in relatie tot de publieke waarden van de vrijheid van meningsuiting en non-discriminatie. Er is een risico dat uitingen ten onrechte als onrechtmatig of als onwenselijk worden gekwalificeerd en aangebracht en dat werkelijk onrechtmatige dan wel onwenselijke uitingen niet als zodanig worden herkend. Daaraan draagt bij dat algoritmen niet altijd goed in staat zullen zijn om zich rekenschap te geven van de betekenis van een uiting en de context waarbinnen een uiting wordt gedaan. Deze risico's zullen tot op zekere hoogte ook gelden ten aanzien van het modereren van andere typen onrechtmatige of ongewenste content. Naarmate een verwijderbeslissing een

[116] Gillespie 2018.

genuanceerde juridische beoordeling vraagt, is de kans groter dat een modereeralgoritme onterecht bepaalde content wel of niet aanbrengt.

Ook het gevaar van vooroordelen speelt daarbij een rol. Wetenschappelijk onderzoek naar de prestaties van *hate speech* modereeralgoritmen wijst erop dat uitingen door minderheidsgroepen als gevolg van vooringenomenheden een hogere kans hebben om te worden gedetecteerd. Menselijke moderators moeten uiteindelijk beslissen over aangebrachte content en kunnen onjuiste kwalificaties rechtzetten. De hoge druk waaronder menselijke contentmoderators veelal moeten beslissen en de beperkte instructies die zij krijgen, nuanceren echter de waarde van dat menselijke toezicht.

De publieke waarde van rechtsbescherming komt met name in geding door de gebrekkige inzichtelijkheid van de werking van besluitvormingsprocessen. De informatie die door online platformen zelf naar buiten wordt gebracht is doorgaans niet inzichtelijk genoeg om een goed beeld te kunnen vormen van de contentmoderatiepraktijk als geheel en de rol van algoritmen daarin. Dat bemoeilijkt het opstellen van passend beleid en het beoordelen van de bestendigheid van bestaande regelgeving.[117]

Gebruikers komen, voor zover bekend, niet te weten of een algoritme een rol heeft gespeeld bij de totstandkoming van dat besluit en wat de eventuele bijdrage van het algoritme was. Hierdoor kan de rechtsbescherming van gebruikers jegens private online platformen – die in veel gevallen al beperkt is als gevolg van de voorwaarden die gebruikers contractueel aanvaarden – onder druk komen te staan. Als platformen in de toekomst echter (sommige) content volledig geautomatiseerd of semi-geautomatiseerd modereren, dan kan de AVG mogelijk uitkomst bieden voor het inzichtelijker maken van het modereerproces en de rol die algoritmen daarin spelen.

In regelgeving met betrekking tot het modereren van content in het algemeen en in het bijzonder ten aanzien van *hate speech* is nog maar nauwelijks aandacht voor de gevolgen van het algoritmisch aanbrengen van onrechtmatige content. Hoewel de druk op platformen wordt opgevoerd om proactief te modereren en daarvoor ook technologie in te zetten, is voor de risico's van de inzet van algoritmen, zoals de mogelijke discriminerende effecten, nog onvoldoende aandacht. Menselijk toezicht op contentmoderatie wordt wel aangemoedigd. Maar bij gebrek aan voorwaarden waaraan dat toezicht inhoudelijk moet voldoen, is een modereerpraktijk ontstaan waarin menselijke moderators onder grote druk moeten beslissen, en waarbij eventuele risico's

[117] Speciaal rapporteur voor de bevordering en bescherming van het recht op vrijheid van meningsuiting aan de VN David Kaye benadrukt in dat verband dat transparantie noodzakelijk is bij de automatisering van contentmoderatie vanwege mogelijke risico's ten aanzien van grondrechten. Daarnaast is het volgens hem van wezenlijk belang dat de samenleving en maatschappelijke belangenorganisaties worden betrokken bij de implementatie van zulke geautomatiseerde processen (Kaye 2018, p. 18).

die samenhangen met de inzet van algoritmen niet of onvoldoende door het huidige juridisch kader worden gemitigeerd.