Countering online hate speech: how to adequately protect fundamental rights?

Nave, E.V.R.

Samenvatting (Dutch Summary)

De bestrijding van online haat vanuit fundamenteelrechtelijk perspectief

Dit manuscript onderzoekt juridische benaderingen die in overeenstemming zijn met de mensenrechten om haatzaaiende uitlatingen op online platforms tegen te gaan. Er is een groeiende prevalentie van haatzaaiende uitlatingen op online platforms. Online platforms hebben zelfregulerende beleidsmaatregelen ontwikkeld om haatzaaiende uitlatingen tegen te gaan. Echter, dergelijke private regelgevende kaders blijven vaak ondoorzichtig en ontberen democratische handhavings- en herstelmechanismen. Dit onderzoek bouwt voort op een kritische conceptualisering van online haatzaaiende uitlatingen, gebaseerd op het Europese regelgevende en beleidskader, om juridische wegen te onderzoeken en voor te stellen die de zorgvuldigheidsplichten met betrekking tot mensenrechten (HRDD) van online platforms versterken. Dit met als doel om strafbare online haatzaaiende uitlatingen tegen te gaan en tegelijkertijd fundamentele rechten te waarborgen.

Hoofdstuk 1 legt de basis door de context en sociale relevantie te presenteren, de probleemstelling en onderzoeksvragen te introduceren en de methodologie en reikwijdte uit te leggen. Online social media-platforms kunnen in brede zin worden omschreven als internet-hostingdiensten die door gebruikers gegenereerde inhoud opslaan, beoordelen, promoten of degraderen voor het algemene publiek via groepen, gepersonaliseerde nieuwsfeeds en berichtenapplicaties. Met meer dan de helft van de wereldbevolking als actieve gebruikers van online socialemediaplatforms, en met het gebruik van deze online omgevingen om fundamentele mensenrechten uit te oefenen, zoals vrijheid van meningsuiting, vrijheid van vergadering en vereniging, heeft de verspreide inhoud en de implementatie van moderatiebeleid een steeds grotere impact op wereldschaal. Mensenrechtenactivisten, klokkenluiders, voormalige werknemers van online platforms en zelfs de Verenigde Naties hebben gewaarschuwd dat sommige sociale mediabedrijven met hun bedrijfsmodellen niet alleen hebben gefaald in het verwijderen van haatzaaiende uitlatingen, maar deze zelfs hebben versterkt. Verdere studies waarschuwen ook dat de toename van haatzaaiende uitlatingen in digitale omgevingen kan resulteren in offline haatzaaiende uitlatingen en haatmisdrijven.

Als reactie op deze ontwikkelingen en onder druk van staten, mensenrechtenactivisten en het maatschappelijk middenveld, zijn sommige online plat-

forms begonnen met zelfregulering van haatzaaiende uitlatingen, het delen van gegevens over de prevalentie van haatzaaiende uitlatingen en het opzetten van toezichtsorganen voor beroepsprocedures inzake contentmoderatie. Echter, dergelijke zelfregulerende inspanningen worden vaak bekritiseerd vanwege het niet naleven van mensenrechtennormen. Belangrijke kritiekpunten zijn onder meer: i) de conceptualisering van haatzaaiende uitlatingen die door platforms wordt gehanteerd; ii) de handhavingsmechanismen voor contentmoderatiebeleid; iii) de herstelmechanismen die gebruikers ter beschikking staan om contentmoderatiebeslissingen aan te vechten. In een poging om de reguleringskaders die door bedrijven worden ingesteld om online haatzaaiende uitlatingen tegen te gaan, democratisch te reguleren en te controleren, produceren zowel staten als internationale en regionale organisaties sectorspecifieke juridische en beleidsinstrumenten. Niettemin rijzen er discussies over de effectiviteit en geschiktheid van dergelijke regelgevende kaders bij het bevorderen van respect voor de mensenrechten van mensen die het doelwit zijn van haatzaaiende uitlatingen.

Daarom stelt dit manuscript de volgende vraag: Gebaseerd op een kritische conceptualisering van online haatzaaiende uitlatingen, en meer specifiek van strafbare haatzaaiende uitlatingen binnen het Europese regelgevende en beleidskader, hoe kunnen Europese wetgevers, zowel op het niveau van de Europese Unie als de Raad van Europa, de zorgvuldigheidsplichten met betrekking tot mensenrechten van online platforms verduidelijken om online haatzaaiende uitlatingen tegen te gaan en tegelijkertijd fundamentele rechten te waarborgen? De gebruikte methodologie is drieledig: doctrinair; vergelijkend; en interdisciplinair. Doctrinair juridisch onderzoek van juridische en beleidskaders die van toepassing zijn op online haatzaaierij probeert bestaande juridische normen, mazen en mogelijke toekomstige juridische wegen te verduidelijken. Vergelijkende juridische analyse wordt gebruikt om de afstemming, of het gebrek daaraan, tussen de normen die door onlineplatforms worden aangenomen via hun servicevoorwaarden en Europese mensenrechtennormen te onderzoeken. Interdisciplinair onderzoek combineert bevindingen uit juridische, sociologische en digitale technologiestudies om zorgen te systematiseren en contentmoderatiepraktijken voor te stellen die geschikt zijn om online haatzaaierij tegen te gaan en die voldoen aan de mensenrechten.

Hoofdstuk 2 stelt een nieuwe juridische conceptualisering van haatzaaiende uitlatingen in de Europese context voor. Huidige kaders missen een gestandaardiseerde benadering van de conceptualisering van haatzaaiende uitlatingen. Sommige conceptualiseringen zijn te breed, en andere zijn onvoldoende inclusief; te breed omdat ze leiden tot het verwijderen van legale content (bijv. verwijderingstools die legale content verwijderen die is geplaatst door gemarginaliseerde gemeenschappen), en onvoldoende inclusief omdat de context van posts door taalkundige minderheden vaak wordt genegeerd. Dit hoofdstuk analyseert het Europese regelgevingskader door de lens van de eerste juridische conceptualiseringen van haatzaaiende uitlatingen die voortvloeien uit de

kritische (rassen)theorie en (zwarte) feministische intersectionaliteitstheorie. Er zijn twee belangrijke bevindingen uit dit hoofdstuk. Ten eerste suggereert dit hoofdstuk dat het Europese regelgevingskader expliciet de conceptualisering van haatzaaiende uitlatingen door kritische rechtsgeleerden moet erkennen als uitingen die bedoeld zijn om historische of systematische onderdrukking in stand te houden. Ten tweede bepleit dit hoofdstuk dat de conceptualisering van haatzaaiende uitlatingen in de Europese context alleen juridische samenhang kan bereiken wanneer alle Europese regelgevingsinstrumenten uitdrukkelijk rekening houden met de intersectionaliteit van systemen van onderdrukking.

Hoofdstuk 3 bevordert specifieke preventieve HRDD-verantwoordelijkheden die van toepassing zijn op onlineplatforms die online haatzaaiende uitlatingen tegengaan. Er wordt meer aandacht besteed aan de HRDD-verantwoordelijkheden van bedrijven die van toepassing zijn op onlineplatforms om online haatzaaiende uitlatingen tegen te gaan. Op het niveau van de Europese Unie reguleren sectoroverschrijdende initiatieven de rechten van gemarginaliseerde groepen en stellen HRDD-verantwoordelijkheden vast voor onlineplatforms om online haatzaaiende uitlatingen snel te identificeren, voorkomen, beperken, verhelpen en verwijderen. Niettemin heeft het HRDD-kader dat van toepassing is op online haatzaaiende uitlatingen zich vooral gericht op de verantwoordelijkheden van de platforms gedurende hun hele bedrijfsvoering – richtlijnen met betrekking tot HRDD-vereisten met betrekking tot de regulering van haatzaaiende uitlatingen in de Servicevoorwaarden van de platforms ontbreken. Dit hoofdstuk gebruikt een conceptualisering van criminele haatzaaiende uitlatingen zoals uitgelegd in de Aanbeveling CM/Rec(2022)16, paragraaf 11, van het Comité van Ministers van de Raad van Europa om specifieke HRDD-verantwoordelijkheden te ontwikkelen. Dit onderzoek omvat een empirische kwalitatieve analyse van drie casestudies: Facebook (Meta Platforms, Inc.), X Corp. (voorheen Twitter, Inc.) en YouTube. Deze empirische analyse beoordeelt de naleving van de Servicevoorwaarden van de platforms met de conceptualisering van criminele haatzaaiende uitlatingen in CM/Rec(2022)16. Dit hoofdstuk beweert dat onlineplatforms, als onderdeel van de opkomende preventieve HRDD-verantwoordelijkheden binnen Europa, de rechten van historisch onderdrukte gemeenschappen moeten respecteren door hun Servicevoorwaarden af ??te stemmen op de conceptualisering van criminele haatzaaiende uitlatingen in Europese mensenrechtennormen.

Hoofdstuk 4 stelt een nieuwe wettelijke minimumstandaard voor die de verantwoordelijkheden van onlineplatforms die E2EE-diensten aanbieden op het gebied van mensenrechten uitbreidt om een ??categorie van criminele haatzaaiende uitlatingen te beperken: aanzetten tot geweld. Diensten die door onlineplatforms worden aangenomen, hebben de verspreiding van online haatzaaiende uitlatingen mogelijk gemaakt. Met name end-to-end gecodeerde (E2EE) diensten staan ??onder toenemende controle voor het hosten van haatzaaiers. Juristen en wetshandhavers worstelen met het conceptualiseren

van de verantwoordelijkheden van E2EE-diensten om geen haatzaaiende uitlatingen te hosten zonder de rechten van gebruikers op vrijheid van meningsuiting, vereniging, privacy of gegevensbescherming onevenredig te beïnvloeden. Na het vaststellen van het algemene HRDD-kader voor HRDD van bedrijven met kunstmatige intelligentie om criminele haatzaaiende uitlatingen te beperken, gaat dit hoofdstuk dieper in op de digitale technologieën en encryptiefuncties die worden gebruikt voor contentmoderatie in E2EE-diensten. Deze analyse past het HRDD-kader toe in combinatie met homomorfe encryptie, metadata en hashing op geselecteerde criminele haatzaaiende uitlatingen die aanzetten tot geweld. Bovendien verduidelijkt dit hoofdstuk de normen voor samenwerking tussen onlineplatforms en wetshandhaving in de context van aanzetten tot geweld in grote groepschats op E2EE-services. Tot slot stelt hoofdstuk 4 een nieuwe wettelijke norm voor die de corporate HRDD van onlineplatforms die E2EE-services aanbieden uitbreidt door middel van regulering en toepassing van metadata, hashing en homomorfe encryptie om aanzetten tot geweld in grote groepen op E2EE-services te verstoren.

Hoofdstuk 5 stelt een uitgebreid kader voor herstelverantwoordelijkheden voor onlineplatforms voor die criminele haatzaaiende uitlatingen hebben veroorzaakt of ertoe hebben bijgedragen, op basis van het algemene kader voor verantwoordelijkheden van bedrijven voor mensenrechten. Wetgevers hebben bindende juridische kaders ontwikkeld die de due diligence- en aansprakelijkheidsregimes voor mensenrechten van deze platforms verduidelijken om haatzaaiende uitlatingen te identificeren en te voorkomen. Deze juridische kaders verduidelijken echter niet de herstelverantwoordelijkheden van onlineplatforms om mensen te herstellen die schade hebben geleden door criminele haatzaaiende uitlatingen die door de platforms zijn veroorzaakt of waaraan deze hebben bijgedragen. De bijdrage van Meta aan de genocide op de Rohingya in Myanmar wordt geanalyseerd als een van de meest grondig gedocumenteerde gevallen die de maatschappelijke impact laten zien van de verantwoordelijkheden van bedrijven voor mensenrechten van zeer grote onlineplatforms die bijdragen aan de versterking van criminele haatzaaiende uitlatingen.

Dit hoofdstuk onderzoekt de toepassing van het recht op een doeltreffende remedie op gevallen van online haatzaaiende uitlatingen. Dit onderzoek onderzoekt ook de internationale normen voor het recht op remedie voor gevallen van grove schendingen van mensenrechten, waarbij wordt erkend dat sommige elementen van criminele haatzaaiende uitlatingen kunnen worden geclassificeerd als grove schendingen van mensenrechten. Na het verduidelijken van het algemene kader voor corrigerende verantwoordelijkheid van bedrijven, dat betrekking heeft op verantwoordelijkheidswijzen, herstelprocessen en herstelresultaten, verduidelijkt dit hoofdstuk dat het herstelkader van toepassing is op onlineplatforms die criminele haatzaaiende uitlatingen hebben veroorzaakt of daaraan hebben bijgedragen. Dit hoofdstuk benadrukt de noodzaak van en stelt een kader voor corrigerende verantwoordelijkheden van bedrijven op EU-niveau voor, ook voor onlineplatforms die criminele

haatzaaiende uitlatingen hebben veroorzaakt of daaraan hebben bijgedragen. Het voorgestelde kader onderzoekt garanties van niet-herhaling, restitutie en compensatie als geschikte herstelresultaten.

Hoofdstuk 6 presenteert de belangrijkste bevindingen met betrekking tot de probleemstelling en onderzoeksvragen die deze thesis motiveren, doet aanbevelingen en bespreekt gebieden voor toekomstig onderzoek. Dit hoofdstuk doet een uitgebreide reeks aanbevelingen om de verantwoordelijkheden van onlineplatforms op het gebied van mensenrechten van bedrijven te versterken om criminele haatzaaiende uitlatingen tegen te gaan. Deze aanbevelingen zijn gericht aan drie actoren, namelijk wetgevers en beleidsmakers, wetshandhavingsinstanties en onlineplatforms. Over het algemeen kunnen wetgevers en beleidsmakers met meer vertrouwen vertrouwen op het algemene HRDD-kader om preventieve, verzachtende en herstellende verantwoordelijkheden voor onlineplatforms te conceptualiseren om criminele haatzaaiende uitlatingen tegen te gaan. Wetshandhavingsinstanties moeten de oprichting van rapportage- en onderzoekskanalen voor criminele haatzaaiende uitlatingen op onlineplatforms vergemakkelijken. Onlineplatforms moeten zich houden aan het HRDD-kader en duidelijke en transparante processen ontwikkelen om criminele haatzaaiende uitlatingen die via hun diensten worden verspreid, te voorkomen, te verzachten en te herstellen. De regulering van onlineplatforms brengt complexe juridische kwesties met zich mee in verschillende onderzoeksdisciplines, met gevolgen voor een veelheid aan nationale en internationale rechtsgebieden en met betrekking tot talrijke belanghebbenden, waaronder onlineplatforms, overheidsinstanties en individuen. Gezien de voortdurend veranderende aard van de diensten die onlineplatformen aanbieden, is het van belang dat toekomstig onderzoek naar maatregelen om online haatzaaierij tegen te gaan, sterker interdisciplinair onderzoek vereist dat zich richt op mensenrechten.