Aanpak van desinformatie met trusted flaggers



Met financiële ondersteuning van:

SIDN fonds

Otto Volgenant en Karlijn Han
30 september 2026

© Boekx Advocaten
Overname uit dit rapport is toegestaan met bronvermelding

Het onderzoek

1.1 Met financiële ondersteuning van het SIDN Fonds heeft Boekx Advocaten in kaart gebracht welke rol trusted flaggers kunnen spelen bij de aanpak van desinformatie. Boekx Advocaten heeft ook specifiek gekeken naar een potentiële rol voor Offlimits. Daarnaast heeft Boekx Advocaten onderzocht of andere organisaties mogelijk geschikt zijn om desinformatie te melden.

1.2 Trusted flaggers zijn entiteiten die meldingen kunnen maken van illegale inhoud bij online platforms. Online platforms moeten meldingen van trusted flaggers met prioriteit behandelen. Zo dragen trusted flaggers bij aan een snelle en betrouwbare aanpak van illegale content. De ACM wijst trusted flaggers aan om specifieke illegale content te melden. Daarvoor moeten ze aantoonbare expertise hebben. Trusted flaggers kunnen dus alleen berichten met desinformatie melden die illegaal zijn én binnen hun expertisegebied vallen.

1.3 Offlimits kan momenteel als trusted flagger meldingen indienen voor slachtoffers van online grensoverschrijdend gedrag. De rol van Offlimits bij het tegengaan van desinformatie is beperkt. Offlimits kan meldingen indienen voor slachtoffers van lasterlijke berichten of deepfakes die worden verspreid zonder hun toestemming. Verder kan Offlimits misleidende advertenties melden als bestaande personen hierin voorkomen. Offlimits kan ook melding maken van deepfakes zonder transparantielabel op pornografische websites. In de praktijk doet Offlimits al geregeld dit soort meldingen.

1.4 Er zijn mogelijk andere organisaties geschikt om desinformatie binnen hun expertisegebied te melden. Dit zijn bijvoorbeeld organisaties binnen het samenwerkingsverband HEIO en Meld Online Discriminatie.

2. Het probleem: online desinformatie

2.1 Desinformatie bedreigt onze democratische samenleving. Een essentiële pijler in de democratie is het publieke debat, waarin ‘goed geïnformeerde burgers in een vrij en eerlijk politiek proces hun wil kunnen uiten’.[1] Desinformatie doet afbreuk aan de werking van dit publieke debat. Burgers zijn minder goed geïnformeerd door desinformatie, waardoor ze minder weloverwogen keuzes maken. Dat geldt ook bij het stemmen tijdens verkiezingen. Daarnaast doet desinformatie afbreuk aan het vertrouwen van burgers in politieke instellingen en in media.[2]

2.2 Desinformatie is van alle tijden. Met de komst van online platforms is het probleem aanzienlijk toegenomen. Al in 2018 waarschuwde de Europese Commissie dat sociale media desinformatie verspreiden ‘op een schaal en met [een] snelheid als nooit tevoren’.[3] Dat is alleen maar versterkt met de komst van artificiële intelligentie (‘AI’). Tegenwoordig kan iedere burger eenvoudig, snel en goedkoop op grote schaal deepfakes maken en verspreiden.

2.3 Het *Hybrid Election Integrity Observatory* (‘HEIO’) heeft onderzoek gedaan naar de integriteit

[1] Europese Commissie, *Tackling online disinformation: a European Apparoach, Com(2018) 236 final,* p. 1, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236.
[2] Europese Commissie (n 1) p. 1.
[3] Europese Commissie (n 1), p. 1.

van de Tweede Kamerverkiezingen van 2025.[4] Volgens HEIO stonden die verkiezingen aanzienlijk onder druk door desinformatie:

'De Nederlandse verkiezingen verliepen vrij en eerlijk, maar stonden online onder aanzienlijke druk. Online campagnes maakten op grote schaal gebruik van AI gegenereerde content. Verder zagen we gecoördineerde manipulatiecampagnes, het falen van platformmoderatie, en het benutten van nieuwe technieken om desinformatie en haatzaaiende uitingen te verspreiden.'[5]

2.4 HEIO identificeert verschillende soorten desinformatie die werden verspreid. Denk daarbij aan onwaarheden over politici en complottheorieën over verkiezingsfraude. Ook gingen deepfakes rond, zoals deepfakes over geweld door immigranten en over de LHBTIQ+-gemeenschap.[6] Problematisch is dat politici soms ook zelf deepfakes verspreidden als onderdeel van hun online campagne.[7]

2.5 Ook waren er grootschalige desinformatiecampagnes vanuit het buitenland. RTL bracht bijvoorbeeld een netwerk van nepaccounts aan het licht die vanuit Nigeria, Ghana en Ivoorkust extreem gedachtegoed verspreidden.[8] Deze campagnes werden mogelijk gefinancierd vanuit Rusland.[9]

2.6 Volgens HEIO was maar 1,3% van de onderzochte content gegenereerd met AI. Desondanks is er grote reden tot zorg, aldus HEIO: juist AI-gegenereerde content kreeg een groot bereik en werd doelgericht verspreid door kwaadwillende actoren. Daarnaast was slechts 32% van alle deepfakes voorzien van een transparantielabel.[10] Ook politieke reclames waren nauwelijks gelabeld.[11]

2.7 Naast desinformatie die direct inmengt in onze democratie zijn er andere soorten desinformatie die schadelijk zijn voor onze samenleving. Zo verspreiden criminelen doelbewust deepfakes om personen op te lichten, bijvoorbeeld deepfakes van influencers die zogenaamd advies geven over crypto-investeringen. Criminelen proberen hiermee gebruikers ertoe te bewegen om crypto-tokens over te maken. Dit worden '*phishing scams*' genoemd.

2.8 Verder gaan deepfakes rond waarin bekende dokters schadelijke gezondheidsclaims doen, bijvoorbeeld van oud-huisarts Ted van Essen. Hij raadde diabetespatiënten zogenaamd aan om te stoppen met het spuiten van insuline. Als patiënten daarmee stoppen kunnen ze in het

---

[4] HEIO, *Hybrid Election Integrity Observatory: Final Report – Monitoring the Dutch Parliamentary Elections*, *29 oktober 2025*, januari 2026, Den Haag: https://www.heio.nl/wp-content/uploads/2026/01/260115-HEIO-Final-Report.pdf.
[5] HEIO (n 4), p 6.
[6] HEIO (n 4), p. 42.
[7] HEIO (n 4), p. 27.
[8] J. Bunskoek e.a., 'Trollenlegers uit buitenland versterkten politieke en opruiende berichten rond verkiezingen', 25 november 2025, *RTL*: https://www.rtl.nl/nieuws/onderzoek/artikel/5540541/trollen-uit-het-buitenland-maakten-rond-nederlandse-verkiezingen.
[9] HEIO (n 4), p. 34.
[10] HEIO (n 4), p. 27.
[11] HEIO (n 4), p. 45.

ziekenhuis belanden of zelfs overlijden.[12]

Desinformatie: definities

2.9 Omdat desinformatie voorkomt in verschillende vormen, is het lastig om vast te stellen wat desinformatie precies is. Academici gebruiken verschillende definities.[13]

2.10 Humprecht e.a. gebruiken de volgende definitie:

'de verspreiding van grotendeels of geheel (wetenschappelijk) ongefundeerde claims waarvan de verzender weet dat ze niet kloppen. Het doel van het maken en verspreiden kan zijn om geld te verdienen, voor vermaak, om het democratisch proces te verstoren of een combinatie hiervan.'[14]

2.11 Volgens een rapport van het Rathenau Instituut uit maart 2026 komt deze definitie de afgelopen tijd het meest voor in de academische literatuur en het Nederlandse publieke debat.[15]

2.12 Wardle en Derakhshan definiëren desinformatie als volgt in hun rapport *Information Disorder* uit 2017:

'*Information that is false and deliberately created to harm a person, social group, organization or country.'*[16]

2.13 De Europese Unie hanteert de volgende juridische definitie:

'*aantoonbaar foute of misleidende informatie die wordt gecreëerd, gepresenteerd en verspreid voor economisch gewin of om het publiek opzettelijk te bedriegen, en die schade in het publiek domein kan veroorzaken*.'[17]

2.14 In dit rapport staat EU-wetgeving centraal. Wij sluiten daarom aan bij de definitie van de EU.

2.15 Deepfakes zijn een vorm van desinformatie. Net als het begrip desinformatie bestaan er verschillende definities van het begrip 'deepfake'. De AI Act definieert een 'deepfake' als

---

[12]https://eenvandaag.avrotros.nl/artikelen/artsen-waarschuwen-voor-toename-deepfakes-met-foute-medische-informatie-schadelijk-voor-patienten-162709.

[13] Zie voor een overzicht van verschillende definities: J. van Hoboken e.a., *Het juridisch kader voor de verspreiding van desinformatie via internetdiensten en de regulering van politieke advertenties – Eindrapportage*, december 2019, IViR: https://www.ivir.nl/publicaties/download/Rapport_desinformatie_december2019.pdf.

[14] Humprecht e.a., 'Resilience to Online Disinformation: A Framework for Cross-National Comparative Research, 2020, *The International Journal of Press/Politics* 2020, 25(3), p. 493-510, https://doi.org/10.1177/1940161219900126.
researc/168076277c

[15] L. van Eeden e.a., *Scrollend naar de stembus – De rol van aanbevelingsalgoritmen bij inmenging in verkiezingen*, Rathenau Instituut 2026, p. 53: https://www.rathenau.nl/nl/digitalisering/naar-een-nieuwe-verhouding-tot-technologiebedrijven/scrollend-naar-de-stembus.

[16] C. Wardle, & H. Derakhshan, *Information Disorder. Toward an interdisciplinary framework of research and policymaking*, 2017 p. 20. Volgens Wardle en Derakhshan verschilt desinformatie van misinformatie, omdat bij misinformatie de intentie van de verzender geen rol speelt.

[17] Europese Commissie (n 1), par. 3.1.1. Zie voor een vergelijking tussen de verschillende definities: Van Hoboken e.a. (n 13), p. 15-18.

volgt:

'*door AI gegenereerd of gemanipuleerd beeld-, audio- of videomateriaal dat een gelijkenis vertoont met bestaande personen, voorwerpen, plaatsen, entiteiten of gebeurtenissen, en door een persoon ten onrechte voor authentiek of waarheidsgetrouw zou worden aangezien'.*[18]

2.16 Wat ingewikkeld is aan deze definitie, is dat het moet gaan om audio- en beeldmateriaal dat 'ten onrechte voor authentiek of waarheidsgetrouw zou worden aangezien'. Of daarvan sprake is, kan per persoon verschillen.

Waarom is desinformatie ingewikkeld om aan te pakken?

2.17 Desinformatie is complex om aan te pakken. Dat komt doordat initiatieven tegen desinformatie snel raken aan de vrijheid van meningsuiting. De vrijheid van meningsuiting vormt de hoeksteen van de democratische samenleving. Onder de vrijheid van meningsuiting valt het recht om informatie te verspreiden, maar ook het recht om informatie te ontvangen. Deze rechten worden onder andere beschermd door artikel 10 EVRM, artikel 11 EU-Handvest en artikel 7 van de Grondwet.

2.18 Wetgevers moeten de vrijheid van meningsuiting waarborgen. Dit betekent dat wetgevers niet zomaar desinformatie kunnen verbieden. Zo'n wet zou al snel een inbreuk vormen op de vrijheid van meningsuiting.

2.19 Met name uitingen die worden gedaan binnen het publieke debat worden ruim beschermd onder de vrijheid van meningsuiting. Dat geldt ook voor satirische content.[19] Er is daarom weinig ruimte voor de wetgever om dit soort uitingen te reguleren. Tegelijkertijd kunnen dit soort uitingen grote impact hebben op democratische processen als ze onwaarheden bevatten. Desinformatie die schadelijk is voor onze samenleving maar wel is toegestaan wordt daarom vaak '*lawful but harmful*' genoemd.[20]

2.20 Tot slot is het soms lastig om te bepalen of desinformatie wel of niet verboden is. Omdat desinformatie raakt aan de vrijheid van meningsuiting, moet per specifiek geval worden beoordeeld of de uiting wel of niet was toegestaan.

Hoe pakt de Europese Unie desinformatie aan?

2.21 In 2018 publiceerde de Europese Commissie haar strategie om desinformatie te bestrijden.[21] Daaruit volgt dat één van de belangrijkste pijlers bestaat uit het creëren van meer transparantie in de online omgeving.[22]

2.22 Het eerste initiatief tegen desinformatie is de *Code of Practice on Disinformation* uit 2018 ('CoP 2018'). De CoP 2018 is tot stand gekomen door bedrijven binnen de tech sector.

[18] Artikel 3(60) AI Act. Zie de AI Act hier: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.
[19] Zie bijvoorbeeld: EHRM 25 januari 2007, appl. nr. 68354/01 (*Vereinigung Bildender Künstler v. Oostenrijk*).
[20] Van Hoboken e.a. (n 13), p. 32.
[21] https://eur-lex.europa.eu/legal-content/NL/TXT/PDF/?uri=CELEX:52018DC0236.
[22] https://eur-lex.europa.eu/legal-content/NL/TXT/PDF/?uri=CELEX:52018DC0236, zie p. 6.

Daarom is de CoP 2018 een vorm van zelfregulering.[23] De CoP 2018 is in 2022 vervangen door de *Strengthened Code of Practice on Disinformation* ('CoP 2022').[24] De Europese Commissie was betrokken bij het opstellen van deze code. Daarom is de CoP 2022 een vorm van co-regulering.[25]

2.23 Platforms kunnen de CoP 2022 ondertekenen. Daarmee gaan ze akkoord met bepaalde verplichtingen. Zo moeten ondertekenaars desinformatie labelen en samenwerken met Fact-checkers. Onder andere Facebook, Instagram, LinkedIn, TikTok en YouTube hebben de CoP 2022 ondertekend. X was eerst ook ondertekenaar, maar heeft zich in 2023 teruggetrokken.[26]

2.24 De belangrijkste wet voor online platforms is de *Digital Services Act* ('DSA').[27] Het hoofddoel van deze wet is om illegale content te bestrijden.[28] De DSA pakt ook desinformatie aan. Dit doet de DSA bijvoorbeeld via risicobeheerverplichtingen (artikelen 34 en 35 DSA). Op basis hiervan moeten *Very Large Online Platforms* ('VLOPs') systeemrisico's op hun platforms beheren.

2.25 VLOPs zijn platforms met maandelijks meer dan 45 miljoen actieve gebruikers in de Europese Unie.[29] Dit soort platforms kunnen een grote impact hebben op de samenleving vanwege hun omvang.[30] Tegelijkertijd is het complex om vast te stellen welke impact VLOPs precies hebben. Daarom zijn VLOPs verplicht om systeemrisico's op hun platforms te identificeren (artikel 34 DSA) en te mitigeren met gepaste maatregelen (artikel 35 DSA). Bijvoorbeeld Instagram, TikTok en X zijn een VLOP. De Europese Commissie houdt toezicht op de naleving van deze verplichtingen.

2.26 Artikel 34 DSA benoemt bepaalde systeemrisico's, waaronder de verspreiding van illegale content (sub a) en negatieve gevolgen voor het maatschappelijk debat en verkiezingsprocessen (sub c).[31]

2.27 Op basis van die laatste categorie zijn VLOPs verplicht om desinformatie en deepfakes aan te pakken. Dit volgt uit de *Richtsnoeren voor aanbieders van VLOP's en VLOSE's inzake de beperking van systeemrisico's voor verkiezingsprocessen* ('*Richtsnoeren voor verkiezingsprocessen*').[32] De Europese Commissie publiceerde deze richtsnoeren op 8 mei 2024. Richtsnoeren zijn niet-bindend. Desondanks zijn ze van belang voor platforms. In richtsnoeren legt de Europese Commissie namelijk uit hoe zij toezicht houdt op de naleving

[23] https://digital-strategy.ec.europa.eu/en/news/code-practice-disinformation.
[24] https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation.
[25] De CoP 2022 is sinds 13 februari 2025 officieel onderdeel geworden van de DSA. De code is nog steeds niet bindend. Een VLOP kan daarentegen wel de maatregelen van de Code 2022 implementeren om aan te tonen dat deze voldoet aan bepaalde risicobeheerverplichting. https://digital-strategy.ec.europa.eu/en/library/code-conduct-disinformation.
[26]https://dsa-observatory.eu/2023/06/12/twitters-retreat-from-the-code-of-practice-on-disinformation-raises-a-crucial-question-are-dsa-codes-of-conduct-really-voluntary/.
[27] https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng.
[28] De DSA is aangenomen in april 2022 en sinds 17 februari 2024 geheel van kracht.
[29] Artikel 33 lid 1 DSA.
[30] Overweging 76 DSA.
[31] Artikel 34 lid 1 sub a en c DSA.
[32] https://eur-lex.europa.eu/legal-content/NL/TXT/PDF/?uri=OJ:C_202403014.

van bepaalde verplichtingen.

2.28 In de Richtsnoeren voor verkiezingsprocessen geeft de Europese Commissie uitleg over hoe VLOPs systeemrisico's tijdens verkiezingen moeten beheren. De Europese Commissie raadt in dat kader diverse mitigerende maatregelen aan. Zo moeten VLOPs deepfakes, desinformatie en politieke advertenties labelen.[33] Ook moeten VLOPs voorkomen dat hun algoritmes desinformatie aanbevelen.[34] Verder dienen ze in hun algemene voorwaarden specifieke regels op te nemen tegen desinformatie.[35]

3. Trusted flaggers

3.1 Het systeem van trusted flaggers is geregeld in artikel 22 DSA. De Europese Commissie heeft op 29 mei 2026 de *Ontwerprichtsnoeren inzake Trusted Flaggers* ('Ontwerprichtsnoeren') gepubliceerd.[36] Daarin geeft de Europese Commissie verdere uitleg over het systeem van trusted flaggers. De definitieve versie moet nog worden gepubliceerd.

3.2 Voordat de DSA was aangenomen bestond al een vrijwillig systeem van trusted flaggers. Dit systeem bestaat nog steeds. Platforms kunnen op vrijwillige basis samenwerken met partijen als trusted flaggers. YouTube heeft bijvoorbeeld het '*Priority Flagger Program*'.[37] Dit vrijwillige systeem is opgenomen in zelfreguleringsdocumenten, zoals de *Code of Conduct on Countering Illegal Hate Speech* uit 2016 ('Code of Conduct 2016').[38] Dit is een zelfreguleringsdocument dat gericht is op het tegengaan van haatzaaiende content online.[39] Onder meer Facebook, YouTube, X, TikTok, Snapchat en LinkedIn zijn ondertekenaars.

3.3 Het probleem is dat het vrijwillige systeem niet-transparant is. Platforms hoeven niet openbaar te maken welke partijen fungeren als trusted flagger. Dit doet afbreuk aan de onafhankelijkheid van trusted flaggers.[40]

3.4 De DSA pakt dit gebrek aan transparantie aan.

<u>Hoe werkt het trusted flagger-systeem van de DSA?</u>

3.5 Op grond van artikel 16 DSA moeten platforms een zogeheten 'kennisgevings- en actiemechanisme' inrichten. Dit mechanisme moet personen en entiteiten, waaronder trusted

---

[33] Zie bijvoorbeeld overweging 27(g) en 40 Richtsnoeren voor verkiezingsprocessen.
[34] Zie bijvoorbeeld overweging 27(c) sub ii en 27(e) en 40 Richtsnoeren voor verkiezingsprocessen.
[35] Zie overweging 27(h) Richtsnoeren voor verkiezingsprocessen.
[36] DG CONNECT DRAFT FOR TARGETED PUBLIC CONSULATION 29 MAY – 10 JULY 2026 COMMUNICATION FROM THE COMMISSION Commission Guidelines on the trusted flagger mechanism under Article 22 of Regulation (EU) 2022/2065: https://digital-strategy.ec.europa.eu/nl/library/draft-commission-guidelines-trusted-flaggers.
[37] https://support.google.com/youtube/answer/7554338?hl=en.
[38]https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en.
[39] De Code of Conduct Illegal Hate Speech 2016 is inmiddels aangesterkt door de *Code of Conduct on Countering Illegal Hate Speech* + ('Code of Conduct+'). Deze is gepubliceerd op 20 januari 2025. Zie hier: https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online.
[40] J. van de Kerkhof, 'Article 22 Digital Services Act: Building trust with trusted flaggers', 31 maart 2025, *Internet Policy Review:* https://doi.org/10.14763/2025.1.1828.

flaggers, in staat stellen om meldingen in te dienen van illegale content.[41] Platforms zijn op basis van artikel 22 DSA verplicht om meldingen van trusted flaggers 'met prioriteit en onverwijld' af te handelen. Dit geldt voor alle platforms, met uitzondering van platforms van MKB's.[42]

3.6 Trusted flaggers mogen melding maken van content die verboden is op basis van de wet. Content die wettelijk gezien is toegestaan maar verboden is op basis van de algemene voorwaarden van een platform mogen trusted flagger niet melden.[43]

3.7 Het is belangrijk dat trusted flaggers hun meldingen voldoende en nauwkeurig motiveren.[44] Een platform is alleen verplicht om content te verwijderen als uit de melding duidelijk volgt dat sprake is van illegale content, zonder dat het platform zelf een gedetailleerde juridische analyse hoeft uit te voeren.[45]

3.8 Als een platform een melding heeft ontvangen, moet het platform beoordelen of inderdaad sprake is van illegale content. Wanneer dit het geval is, moet het platform de content 'prompt' verwijderen. Anders riskeert het platform aansprakelijkheid.[46] Soms kan een platform ook volstaan met een andere maatregel dan verwijdering. Als een deepfake bijvoorbeeld verboden is omdat deze niet gelabeld is, kan een platform ook ervoor kiezen om een label toe te voegen om aansprakelijkheid te voorkomen.[47]

3.9 Of sprake is van illegale inhoud moet worden beoordeeld aan de hand van nationale wetgeving per lidstaat. De DSA bepaalt dus niet welke content verboden is.[48] In principe hoeven platforms alleen in de lidstaat waar de content verboden is de content te verwijderen. In de praktijk verwijderen grote platforms content vaak wereldwijd vanwege hun grensoverschrijdende karakter.

3.10 Er is geen vaste termijn waarbinnen platforms meldingen moeten afhandelen. Hoe snel een platform een melding moet verwerken is afhankelijk van verschillende factoren, zoals het type illegale inhoud en de kwaliteit van de melding.[49] Uit de DSA volgt dat platforms meldingen van niet-consensueel seksueel beeldmateriaal extra snel moeten afhandelen.[50]

3.11 Dat geldt ook voor meldingen over haatzaaiende content. In de DSA staat dat hiervoor de 'benchmark' van 24 uur geldt.[51] Die termijn is afkomstig uit de Code of Conduct 2016.[52] Ondertekenaars van deze code zijn verplicht om de meerderheid ('majority') van de

[41] Artikel 16 DSA.
[42] Artikel 19 DSA.
[43] Overweging 125 Ontwerprichtsnoeren.
[44] Artikel 16 lid 2 DSA.
[45] Overweging 77 DSA.
[46] Artikel 6 DSA.
[47] Artikel 3 sub t DSA.
[48] Artikel 3 sub h DSA.
[49] Artikel 62 DSA.
[50] Overweging 87.
[51] Overweging 87.
[52] https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en. Zie p. 2: '*The IT Companies to review the majority of valid notification for removal of illegal hate speech in less than 24 hours and remove or disable access to such content, if necessary*.'

meldingen over haatzaaiende berichten binnen 24 uur af te handelen.[53]

Toekenning trusted flagger-status

3.12 De 'nationale digitaledienstencoördinator' is bevoegd om een entiteit de trusted flagger-status toe te kennen.[54] In Nederland is dit de Autoriteit Consument en Markt ('ACM'). Om de status te krijgen, moeten entiteiten een verzoek indienen. Het kan gaan om publieke entiteiten, semipublieke entiteiten, NGO's en private entiteiten zolang ze geen collectief belang vertegenwoordigen.[55] In Nederland zijn op dit moment vier trusted flaggers aangewezen: Stichting Brein, Meld Online Discriminatie, Stichting React en Offlimits.[56]

3.13 Entiteiten kunnen uitsluitend de status verkrijgen om illegale content te melden binnen hun specifieke expertisegebied. Entiteiten moeten beschikken over aantoonbare 'bijzondere deskundigheid en bekwaamheid'.[57]

3.14 Artikel 22 lid 2 DSA noemt alle vereisten waaraan entiteiten moeten voldoen. De entiteit moet i) specifieke expertise en bevoegdheid hebben voor het opsporen, identificeren en melden van illegale inhoud, ii) onafhankelijk zijn van online platforms, en iii) in staat zijn om activiteiten zorgvuldig, nauwkeurig en objectief uit te voeren.[58]

3.15 Een trusted flagger kan ook de bevoegdheid krijgen om meldingen te maken van illegale content in andere lidstaten.[59] Dit is mogelijkerwijs alleen mogelijk als de trusted flagger expertise heeft opgebouwd in het identificeren van illegale content in die lidstaat.

3.16 In principe verkrijgt een entiteit de status voor onbepaalde tijd.[60] De digitaledienstencoördinator is bevoegd om onder bepaalde omstandigheden de trusted flagger-status in te trekken. Dat kan ten eerste wanneer een trusted flagger een groot aantal onnauwkeurige of ongemotiveerde meldingen heeft ingediend.[61] Ook kan dit wanneer een entiteit niet meer voldoet aan de voorwaarden van artikel 22 lid 2 DSA, zoals de onafhankelijkheid en objectiviteit.[62]

Identificatie systeemrisico's

3.17 Eenmaal per jaar moeten trusted flaggers een rapport uitbrengen over de meldingen die ze hebben gedaan en de acties die platforms naar aanleiding daarvan hebben genomen. Dit rapport moeten ze delen met de ACM. De ACM deelt deze rapporten met de Europese

[53] De Code of Conduct+ heeft een systeem van 'Monitoring Reporters' die controleren of platforms daadwerkelijk binnen 24 uur 50% van de meldingen afhandelen. Trusted flaggers kunnen zich aanmelden als Monitoring Reporter. Zie para. 2.3: https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online.
[54] Artikel 22 lid 2 DSA.
[55] Overweging 29 Ontwerprichtsnoeren.
[56] Zie het DSA Jaarverslag 2025 van de ACM: https://www.acm.nl/system/files/documents/dsa-jaarverslag-2025.pdf. De aanwijzingsbesluiten zijn te vinden op de website van de ACM.
[57] Overweging 61 DSA.
[58] Artikel 22 (2) DSA.
[59] Overweging 87 sub d Ontwerprichtsnoeren.
[60] Overweging 56.
[61] Artikel 22 lid 6 DSA.
[62] Artikel 22 lid 7 DSA.

Commissie en digitaledienstenraad van de EU.[63]

3.18 Een trusted flagger kan via dit rapport in kaart brengen wat voor soort illegale content veel voorkomt op een platform. Zo draagt een trusted flagger bij aan de identificatie van een systeemrisico (artikel 34 DSA). Dat kan een verplichting creëren voor een VLOP om mitigerende maatregelen te treffen (artikel 35 DSA).

3.19 Bij het opstellen van dit rapport moeten trusted flaggers rekeninghouden met de taxonomie van illegale content van de Europese Unie. De taxonomie is te vinden in Bijlage II van de Uitvoeringsverordening (EU) 2024/2835.[64] Daarin staan verschillende categorieën illegale content opgenomen, variërend van cybergeweld en schending op bedrijfsgeheimen tot illegale organisaties.[65] Er staan ook categorieën in die gerelateerd zijn aan desinformatie, zoals 'gendergerelateerde desinformatie' (categorie 4d) en 'misinformatie, desinformatie, buitenlandse informatiemanipulatie en inmenging' (categorie 8a).[66]

4. Hebben trusted flaggers een rol bij het melden van desinformatie?

4.1 Op basis van de huidige regelgeving mogen trusted flaggers desinformatie melden die verboden is én binnen hun expertisegebied valt.

4.2 Er is in Nederland geen regelgeving die desinformatie *an sich* verbiedt. Dat zou in strijd zijn met de vrijheid van meningsuiting (hoofdstuk 1). Toch kan in bepaalde gevallen desinformatie verboden zijn. Dat is bijvoorbeeld het geval als desinformatie aanzet tot haat of geweld. In het volgende hoofdstuk bespreken we enkele voorbeelden. Trusted flaggers kunnen dit soort desinformatie melden.

Zijn er trusted flaggers die desinformatie melden?

4.3 In de Europese Unie zijn momenteel vier entiteiten aangewezen als trusted flagger om desinformatie te melden. Dit zijn Factreview, Debunk.org, Instituto Universitário de Lisboa ('ISCTE') en Repsense.[67] Factreview en Debunk.org hebben hun eerste transparantierapport gepubliceerd. Daaruit volgt dat zij desinformatie melden die verboden is.

4.4 Factreview is een NGO uit Griekenland. Op 4 november 2024 werd Factreview aangewezen als trusted flagger. Factreview houdt zich met name bezig met het melden van advertenties met schadelijke gezondheidsclaims en phishing scams.[68]

4.5 Debunk.org is een NGO uit Vilnius en kreeg de trusted flagger-status op 23 maart 2025.

---

[63] Artikel 22 lid 3 DSA.
[64] https://eur-lex.europa.eu/legal-content/NL/TXT/PDF/?uri=OJ:L_202402835.
[65] Uitvoeringsverordening (EU) 2024/2835, zie p. 30-34.
[66] Trusted flaggers en platforms moeten bij het indienen en afhandelen van meldingen rekening houden met de taxonomie. Ook moeten de nationale digitaledienstencoördinatoren rekening houden met de taxonomie bij het toekennen van de trusted flagger-status. Zie: Overwegingen 35, 75, 95(a) en 95(c) Ontwerprichtsnoeren.
[67] In totaal zijn er in de Europese Unie meer dan 70 Trusted Flaggers aangewezen. De Europese Unie moedigt een diversiteit aan Trusted Flaggers aan (overweging 60 DSA). Een overzicht van de aangewezen trusted flaggers is hier te vinden: https://digital-strategy.ec.europa.eu/nl/policies/trusted-flaggers-under-dsa.
[68] factreview.gr/2025/06/05/trusted-flagger-report-1-en; https://factreview.gr/2026/03/03/trusted-flagger-report-2-en/.

Deze NGO is onder andere actief in de Baltische staten, Polen, Georgië en Montenegro. Debunk.org richt zich op advertenties met schadelijke gezondheidsclaims.[69] Daarnaast identificeert Debunk.org grootschalige desinformatiecampagnes. In 2025 ontdekte Debunk.org een netwerk van 1.100 Facebook- en Instagram-accounts. Deze accounts waren zogenaamd van personen van Interpol, Europol en andere EU-instellingen die terugbetaling beloofden vanwege fraude. Het ging om meer dan 50.000 misleidende advertenties. Debunk.org zorgde ervoor dat 874 accounts werden verwijderd, samen met 29.156 advertenties.[70]

4.6 ISCTE is een openbare universiteit in Lissabon. Op 6 februari 2026 ontving ISCTE de trusted flagger-status. Naar verwachting brengt ISCTE in 2027 het eerste transparantierapport uit. Uit de toekenning volgt dat ISCTE zich bezighoudt met het identificeren van desinformatie tijdens verkiezingen. Ook speelde ISCTE een rol bij het tegengaan van desinformatie tijdens Covid-19.[71]

4.7 Repsense, een bedrijf in Litouwen, houdt zich bezig met het identificeren van online desinformatiecampagnes.[72] Op 8 september 2026 heeft Repsense de trusted-flagger status ontvangen.

5. Desinformatie die verboden is in Nederland

5.1 Hieronder worden enkele voorbeelden gegeven van desinformatie die verboden is op basis van nationale en EU-regelgeving. Dit is geen uitputtend overzicht. Ook kan het in sommige gevallen lastig zijn om te bepalen of een uiting daadwerkelijk verboden is. Bij elke uiting moet zorgvuldig rekening worden gehouden met de vrijheid van meningsuiting.

Reclames

5.2 (Politieke) reclames kunnen grote invloed hebben op democratische processen. In de Europese Unie zijn reclames sterk gereguleerd.

AI-gegenereerde reclames zonder label

5.3 Op basis van artikel 50 lid 4 AI Act moeten personen die deepfakes genereren voor commerciële doeleinden deze deepfakes labelen.[73] AI gegenereerde reclames op online platforms moeten daarom gelabeld zijn. Anders zijn ze verboden. Deze verplichting geldt voor alle deepfakes die na 2 augustus 2026 worden verspreid.

5.4 Dit betekent dat politici en politieke partijen die deepfakes verspreiden als onderdeel van hun campagne de deepfakes moeten labelen. De verplichting geldt niet voor de ‘gewone’ burger die in zijn of haar vrije tijd een deepfake maakt en online verspreidt.

[69] https://www.debunk.org/debunk-org-trusted-flagger-2025-report.
[70] https://www.debunk.org/debunk-org-trusted-flagger-2025-report.
[71] https://www.anacom.pt/render.jsp?contentId=1820133.
[72] https://repsense.io/case-studies.
[73] https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

Reclames die niet herkenbaar zijn als reclame

5.5 Een reclame moet herkenbaar zijn als reclame (artikel 6:193b BW en artikel 11 Nederlandse Reclame Code). Dat kan bijvoorbeeld via een label. Reclames die niet als zodanig herkenbaar zijn, zijn illegaal.

5.6 Voor politieke reclames gelden extra strenge transparantie-eisen. Politieke reclames moeten voorzien zijn van een uitgebreide transparantieverklaring. Dit volgt uit artikel 11 van de *EU Verordening betreffende transparantie en gerichte politieke reclame* ('Verordening betreffende politieke reclame').[74] In deze verklaring moet bijvoorbeeld staan wie de opdrachtgever is en met welke verkiezing de reclame verband houdt.[75]

5.7 Er zijn twee categorieën 'politieke reclames'. Ten eerste zijn uitingen die politici en politieke partijen doen binnen een politieke campagne een politieke reclame.[76]

5.8 De tweede categorie omvat uitingen die 'bedoeld' zijn om verkiezingen te beïnvloeden. Dit is een vage categorie. In ieder geval valt desinformatie die wordt verspreid binnen een grootschalige (buitenlandse) desinformatie campagne hieronder. Dit soort berichten zijn dus verboden als ze geen label hebben.

Misleidende advertenties

5.9 Misleidende advertenties zijn verboden op grond van artikel 6:194b BW. Dit zijn bijvoorbeeld advertenties met schadelijke gezondheidsclaims en phishing scams.

Digital Fairness Act

5.10 Er komen nieuwe regels ten aanzien van online advertenties met de aangekondigde *EU Digital Fairness Act* ('DFA'). Het doel van deze wet is om oneerlijke online handelspraktijken aan te pakken. Er komen bijvoorbeeld nieuwe regels voor influencer marketing. Het wetsvoorstel voor de DFA wordt naar verwachting gepubliceerd in Q4 van 2026.[77]

<u>Deepfakes</u>

Deepfakes van personen zonder toestemming

5.11 Het is in het algemeen niet toegestaan om deepfakes van personen te verspreiden zonder toestemming. Maar er zijn uitzonderingen.

5.12 Op grond van artikel 21 van de Auteurswet ('Aw') heeft eenieder het recht om zich te verzetten tegen gebruik van zijn portret als de geportretteerde daarvoor een redelijk belang heeft. Een redelijk belang kan bijvoorbeeld liggen in de bescherming van de privacy of de eer en goede naam van de geportretteerde.

---

[74] https://eur-lex.europa.eu/legal-content/NL/TXT/PDF/?uri=OJ:L_202400900.
[75] Zie artikel 12 lid 1 Verordening betreffende politieke reclame.
[76] Artikel 3 lid 2 sub a Verordening betreffende politieke reclame.
[77] https://www.europarl.europa.eu/legislative-train/theme-protecting-our-democracy-upholding-our-values/file-digital-fairness-act.

5.13 Ook is de Algemene Verordening Gegevensbescherming ('AVG') van toepassing op deepfakes van personen. Bij de verspreiding van dit soort deepfakes worden persoonsgegevens verwerkt. Daarvoor is een verwerkingsgrondslag nodig, zoals toestemming of een gerechtvaardigd belang. Daarnaast moet de betrokkene over de deepfake geïnformeerd worden. Dit maakt deepfakes zonder toestemming al snel strijdig met de privacyregels.

5.14 Bij deepfakes over politici is het vaak lastig om te bepalen of ze zijn toegestaan. Dit soort deepfakes raken snel aan de vrijheid van meningsuiting, omdat ze kunnen worden gezien als een bijdrage aan het maatschappelijke debat. Hetzelfde geldt voor satirische deepfakes.

Deepfakes verspreid door bedrijven

5.15 Bedrijven die deepfakes verspreiden voor commerciële doeleinden zijn verplicht om deepfakes te labelen op basis van artikel 50 lid 4 AI Act. Dit geldt bijvoorbeeld voor aanbieders van websites met pornografische content.

Desinformatie die aanzet tot haat, geweld of discriminatie

5.16 Desinformatie kan aanzetten tot haat, geweld of discriminatie. Denk aan onware nieuwsberichten over immigranten die zogenaamd geweld plegen of discriminerende berichten over de LHBTIQ+-gemeenschap. Uitingen die aanzetten tot haat, geweld of discriminatie zijn verboden op grond van artikel 137d Wetboek van Strafrecht.

5.17 De Europese Commissie heeft de aanpak van online haat hoog op de agenda staan. De Europese Commissie moedigt VLOPs expliciet aan om zich tijdens verkiezingen in te zetten tegen racisme, gendergerelateerde desinformatie, genderelateerd geweld en buitenlandse inmenging gericht op de LHBTIQ+-gemeenschap. Dit staat in de Richtsnoeren voor verkiezingsprocessen.[78]

Laster

5.18 Tot slot zijn lasterlijke berichten verboden op grond van artikel 262 van het Wetboek van Strafrecht. Bij laster beschadigt iemand opzettelijk de reputatie van een ander door onwaarheden te verspreiden, terwijl de verspreider weet dat deze feiten niet kloppen.[79] Het gaat bij laster om concrete beschuldigingen.

6. Desinformatie: mogelijke rol voor Offlimits?

6.1 Het statutaire doel van Offlimits is het voorkomen en bestrijden van online (seksueel) grensoverschrijdend gedrag en misbruik:

[78] Zie Overweging 34 Richtsnoeren voor verkiezingsprocessen. Een ander belangrijk initiatief tegen online haat is de *Richtlijn Audiovisuele Media.* Deze richtlijn is onder andere bedoeld om minderjarigen te beschermen tegen schadelijke online inhoud. https://eur-lex.europa.eu/eli/dir/2010/13/oj/nld. De Richtlijn Audiovisuele Mediadiensten is in 2018 herzien: https://eur-lex.europa.eu/legal-content/NL/ALL/?uri=CELEX:32018L1808.

[79] Artikel 262 Wetboek van Strafrecht: '*Hij die het misdrijf van smaad of smaadschrift pleegt, wetende dat het te last gelegde feit in strijd met de waarheid is, wordt, als schuldig aan laster, gestraft met gevangenisstraf van ten hoogste twee jaren of geldboete van de vierde categorie.'*

'het voorkomen en bestrijden van online grensoverschrijdend gedrag en misbruik in het algemeen en het voorkomen en bestrijden van online seksueel (kinder)misbruik en seksuele (kinder)uitbuiting in het bijzonder, en voorts al hetgeen met één en ander rechtstreeks of zijdelings verband houdt of daartoe bevorderlijk kan zijn, alles in de meest ruime zin van het woord.'

6.2 Offlimits werkt al sinds 1995 op dit gebied en maakt al jarenlang meldingen bij grote platforms. Op basis van deze expertise en ervaring heeft Offlimits een aanvraag ingediend voor het verkrijgen van de trusted flagger-status.

6.3 Op 12 juli 2025 heeft Offlimits de trusted flagger-status toegekend gekregen voor het melden van online (seksueel) grensoverschrijdend gedrag:

'De ACM kent de status van betrouwbare flagger toe aan Stichting Offlimits, op grond van artikel 22, tweede lid, van de DSA, voor het identificeren en melden van online (seksueel) grensoverschrijdend gedrag, waaronder zonder toestemming verspreide naaktbeelden, chantage met naaktbeelden, online grooming, doxing, identiteitsfraude, cyberstalking of cyberpesten en voor het identificeren en melden van online beeldmateriaal van seksueel kindermisbruik'.[80]

6.4 Offlimits kan op basis hiervan binnen het stelsel van de DSA als trusted flagger meldingen indienen voor slachtoffers van online grensoverschrijdend gedrag. De rol van Offlimits bij het tegengaan van desinformatie is momenteel beperkt. Offlimits kan meldingen indienen voor slachtoffers van lasterlijke berichten of deepfakes die worden verspreid zonder hun toestemming. Verder kan Offlimits misleidende advertenties melden als bestaande personen hierin voorkomen. Offlimits kan ook melding maken van deepfakes zonder transparantielabel op pornografische websites. In de praktijk doet Offlimits al geregeld dit soort meldingen.

7. Alternatieve organisaties

7.1 Andere organisaties zijn mogelijk geschikt om niet-toegestane desinformatie binnen hun expertisegebied te melden.

<u>Meld Online Discriminatie</u>

7.2 Meld Online Discriminatie ('MOD') is het meldpunt ter bestrijding van online discriminatie. Dit meldpunt is opgezet door Stichting NL Confidential.[81]

7.3 MOD heeft al ruim 10 jaar ervaring met het ontvangen, beoordelen en het doen van meldingen bij platforms. MOD richt zich op:

'*openbare online haatzaaiende content die zowel strafbaar als discriminerend is, al dan niet in combinatie met het oproepen tot geweld*.'[82]

7.4 Op 29 september 2025 heeft de ACM MOD de trusted flagger-status verleend. MOD is

[80] https://www.acm.nl/system/files/documents/toekenning-status-betrouwbare-flagger-aan-stichting-offlimits.pdf.
[81] https://www.acm.nl/system/files/documents/openbare-versie-aanwijzingsbesluit-meld-online-discriminatie.pdf.
[82] https://www.acm.nl/system/files/documents/openbare-versie-aanwijzingsbesluit-meld-online-discriminatie.pdf.

bevoegd om meldingen te maken voor:

'het identificeren en melden van openbare online haatzaaiende content die zowel strafbaar als discriminerend is, al dan niet in combinatie met het oproepen tot geweld'[83]

7.5 Dit betekent dat MOD zich kan inzetten om desinformatie te melden die aanzet tot haat, geweld en discriminatie.[84]

HEIO

7.6 HEIO is een samenwerkingsverband van experts uit wetenschappelijke en maatschappelijke organisaties. Samen zetten zij zich in voor eerlijke verkiezingen. Onder andere monitort HEIO online verkiezingscampagnes.[85]

7.7 HEIO bestaat uit Post-X Society, Trollrensics, AI Forensics en de Universiteit van Amsterdam in samenwerking met CampAIgn Tracker.[86] Mogelijk komt één of meer organisaties in aanmerking voor de trusted flagger-status.

7.8 Dat geldt op het eerste gezicht met name voor Trollrensics en de Universiteit van Amsterdam. Trollrensics is een Nederlandse B.V. die desinformatie campagnes op online platforms identificeert.[87] De Universiteit van Amsterdam identificeert en onderzoekt samen met onderzoeksinitiatief CampAIgn Tracker het gebruik van AI in het online politieke debat.[88]

[83] https://www.acm.nl/system/files/documents/openbare-versie-aanwijzingsbesluit-meld-online-discriminatie.pdf.
[84] MOD kan zich ook aanmelden als 'Monitoring Reporter' op grond van de Code of Conduct+. Zie voetnoot 53.
[85] https://www.heio.nl/.
[86] https://www.heio.nl/team/.
[87] https://www.trollrensics.com/.
[88]https://gr26.campaigntracker.nl/#info.
https://www.heio.nl/team/;https://www.uva.nl/content/nieuws/nieuwsberichten/2026/03/ai-duikt-overal-op-in-verkiezingscampagnes-maar-dit-wordt-zelden-vermeld.html.

Bijlage: Overzicht relevante regelgeving over trusted flaggers

EU-wetgeving

VERORDENING (EU) 2022/2065 VAN HET EUROPEES PARLEMENT EN DE RAAD van 19 oktober 2022 betreffende een eengemaakte markt voor digitale diensten en tot wijziging van Richtlijn 2000/31/EG *(digitaledienstenverordening)* (https://eur-lex.europa.eu/legal-content/NL/TXT/?uri=CELEX%3A32022R2065).

Uitvoeringsverordening (EU) 2024/2835 van de Commissie van 4 november 2024 tot vaststelling van modellen betreffende de transparantierapportageverplichtingen voor aanbieders van tussenhandelsdiensten en aanbieders van onlineplatforms uit hoofde van Verordening (EU) 2022/2065 van het Europees Parlement en de Raad (https://eur-lex.europa.eu/legal-content/NL/TXT/PDF/?uri=OJ:L_202402835).

EU-beleidsdocumenten

Richtsnoeren van de Commissie voor aanbieders van zeer grote onlineplatforms en zeer grote onlinezoekmachines inzake de beperking van systeemrisico's voor verkiezingsprocessen overeenkomstig artikel 35, lid 3, van Verordening (EU) 2022/2065 (C/2024/3014) (https://eur-lex.europa.eu/legal-content/NL/TXT/PDF/?uri=OJ:C_202403014).

DG CONNECT DRAFT FOR TARGETED PUBLIC CONSULATION 29 MAY – 10 JULY 2026 COMMUNICATION FROM THE COMMISSION Commission Guidelines on the trusted flagger mechanism under Article 22 of Regulation (EU) 2022/2065 (29 mei 2026) (https://digital-strategy.ec.europa.eu/nl/library/draft-commission-guidelines-trusted-flaggers).

Zelf- en co-reguleringsdocumenten

Gedragscode voor de bestrijding van illegale haatzaaiende uitlatingen op het internet (30 juni 2016) (https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en) .

Code of Practice on Disinformation (26 september 2018) (https://digital-strategy.ec.europa.eu/en/news/code-practice-disinformation).

Strengthened Code of Practice on Disinformation (16 juni 2022) (https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation.).

Code of Conduct on Countering Illegal Hate Speech Online + (20 januari 2025) (https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online).