Feit en Fictie in tijden van Generatieve AI

Input voor de NWA-ORC call 2026 voor de route Waardecreatie door verantwoorde AI en Big Data

Jikke van den Ende & Simone van der Burg

15 oktober 2025

waag futurelab

Inleiding

De komst van Generatieve AI (GenAI) en de snelheid waarop er ontwikkelingen plaatsvinden leiden tot grote maatschappelijke uitdagingen. Taal- en beeldmodellen, zoals ChatGPT en Midjourney, maken het voor iedereen mogelijk om op grote schaal synthetische content te genereren. De technologie wordt steeds beter en de output daarmee steeds realistischer. Zo ervaart de Nederlandse bevolking dat het steeds moeilijker wordt om 'echt' van 'nep' te onderscheiden.[1] Mensen maken zich zorgen over de grootschalige verspreiding van onjuiste informatie en het mogelijke misbruik van *deepfakes*.

Content gegenereerd door AI is niet per definitie onwaar, maar het probleem is dat het voor de ontvangers van deze content lastig te beoordelen is. De manier waarop we de waarheid van content kunnen achterhalen werkt niet bij AI-gegenereerde content. Het herleiden van bronnen is lastig, zeker als je weinig kennis hebt over de werking van de technologie. Bovendien wordt het door *Big Tech* bedrijven vaak onmogelijk gemaakt om te achterhalen op welke data de AI getraind wordt en op welke keuzes het model gebaseerd is. Het is een *black box* en dat maakt dat mensen het gevoel hebben grip te verliezen om de wereld om zich heen: kun je de informatie die je te zien krijgt nog wel vertrouwen?

Binnen de route Waardecreatie door verantwoorde Artificial Intelligence en Big Data van de Nationale Wetenschapsagenda (NWA) worden maatschappelijke uitdagingen als deze geadresseerd. Al enkele jaren ondersteunt Waag Futurelab het routemanagement van deze route, bestaande uit Freek Bomhof (TNO) en Tibor Bosse (Radbout Universiteit). Zo is er eerder al maatschappelijke input opgehaald voor de thematische focus van de NWA-ORC call 2024.[2]

Voor het vaststellen van de thematische focus voor de NWA-ORC call 2026 kreeg Waag Futurelab in 2024 de opdracht om een methode te ontwerpen en uit te voeren voor het verder uitdiepen van het thema Feit en Fictie in Tijden van Generatieve AI. De uitkomsten van voorgaand onderzoek dienen daarvoor als fundament. De maatschappelijke relevantie van het thema staat vast, maar door stakeholders uit het veld, zoals journalisten en redacties, technologie ontwikkelaars, beleidsmakers en onderzoekers te betrekken bij de vorming van deze call werkten we toe naar een breed gedragen onderzoeksrichting en vragen voor de

---

[1] https://waag.org/sites/waag/files/2024-04/Eindrapport-Een-maatschappelijke-onderzoeksagenda-voor-AI.pdf

[2] https://waag.org/sites/waag/files/2023-09/20230831_Eindrapport_WaardeCreatieDoorVerantwoordeAI%26BigData.pdf

nieuwe call. Daarnaast onderzochten we ook in breder opzicht hoe de samenleving betrokken kan worden bij agendasetting, call vorming en beoordeling van onderzoeksvoorstellen. Dat onderzoek is nog niet afgerond, dus wordt daarom niet meegenomen in voorliggend rapport.

Proces en resultaten

De maatschappelijke relevantie van het thema Feiten en Fictie in tijden van Generatieve AI komt voort uit eerder onderzoek waarvan de resultaten gepubliceerd zijn in de Maatschappelijke Onderzoeksagenda voor AI.[3] Echter, de achterliggende problematiek is nog behoorlijk breed. Bovendien is deze focus gebaseerd op zorgen en vragen vanuit de samenleving, maar behoeft het nog aanscherping aan de hand van input van huidige stakeholders en experts. Daarom is er in eerste instantie een overzicht gemaakt van wie deze stakeholders en experts zijn. Dit zijn (naast de samenleving) onder andere wetenschappers (zowel gamma als bèta), beleidsmakers, journalisten en technologie ontwikkelaars.

Vervolgens zijn er interviews met acht experts uit verschillende domeinen afgenomen. In deze interviews is het thema Feit en Fictie in tijden van Generatieve AI verder uitgediept, vanuit het perspectief van de geïnterviewde. Deze interviews waren verkennend. Er is gevraagd naar welke maatschappelijke uitdagingen er spelen en wat daar mogelijke oplossingen voor zouden kunnen zijn. Vanuit eigen expertise en ervaringen hebben de geïnterviewden een beeld geschetst van mogelijk waardevolle onderzoeksrichtingen. Een analyse van de transcripten resulteerde in vijf hoofdonderwerpen:

Epistemic Agency

De angst die mensen hebben voor het niet kunnen onderscheiden van "echt" of "nep" uit zich in de *Epistemic Agency* van een individu. Deze term verwijst naar het vermogen om kennis te verwerven, te begrijpen en te gebruiken. Het betreft het vertrouwen in de betrouwbaarheid van informatie en het vermogen om deze informatie vervolgens te integreren in het eigen denken en handelen. *Epistemic Agency* gaat niet over het onderscheid kunnen maken tussen door AI gegenereerde content en "echte" content, maar over het kunnen beoordelen van die content op waarheid en het vertrouwen wat

[3] https://waag.org/nl/article/nederlandse-bevolking-stelt-prioriteiten-voor-onderzoeksagenda-ai/

daarin gesteld wordt. De vraag is wat de impact is van de komst van generatieve AI op de *Epistemic Agency* van een individu.

Misbruik van Generatieve AI

Er kan op allerlei verschillende manieren misbruik gemaakt worden van Generatieve AI. Voorbeelden zijn het genereren van desinformatie of *hate speech,* reputatieschade en (identificatie)fraude. Dit alles leidt tot een verminderd gevoel van veiligheid en vertrouwen. Het gebruik van generatieve AI hoeft niet direct misleidend te zijn. Integendeel, in de (digitale) media kan het als handig hulpmiddel ingezet worden. Maar dan moet het voor de gebruiker wel mogelijk zijn om de waarheid te kunnen achterhalen: hoe ziet het model eruit en waar is het op gebaseerd? Dit vereist een bepaalde mate van kritisch denkvermogen en AI-geletterdheid waar op verschillende manieren (ook technologisch) invulling aan gegeven kan worden.

Tools voor detectie

De effectiviteit van detectie technologie laat nog wat te wensen over. In de praktijk worden tools vaak gezien als onbetrouwbaar en zijn bovendien *biased*. Er vindt een continue wapenwedloop plaats tussen de ontwikkeling van generatieve AI en de technologie voor detectie. De innovatie en het grote geld zit vooral in de innovatie van generatieve AI zelf, waardoor er een achterstand bestaat aan de andere kant die lastig kan worden ingehaald. Daarnaast worden detectie tools vaak ingezet als een hulpmiddel, nooit op opzichzelfstaand. Voorop staat dat er altijd een menselijk beoordeling aan te pas moet komen. Dit maakt het lastig omdat vaak niet duidelijk is wat de achterliggende werking van de tool is en op welke data het gebaseerd is. Detectie alleen is niet voldoende – het gaat er ook om wat het betekent dat iets door AI gegenereerd is en hoe we daar vervolgens mee omgaan.

Onderwijs en AI Literacy

Generatieve AI heeft een grote impact op onze manier van onderwijs geven. Er zijn fundamenteel andere processen nodig om ervoor te zorgen dat we als samenleving niet een achterstand in vaardigheden of kennis krijgen. Dit vraagt om andere lesmethodieken, zoals formatief toetsen in plaats van summatief. Dat betekent dat er minder nadruk wordt gelegd op het eindresultaat, maar op het leerproces zelf. Daarnaast biedt het onderwijs ons de mogelijkheid om het kritisch denkvermogen wat nodig is voor het beoordelen van generatieve content in de media aan te leren. Het is van belang dat er op jonge leeftijd al aandacht wordt besteed aan AI-geletterdheid en mediawijsheid. Daarbij gaat het niet

alleen over het kunnen herkennen van generatieve content, maar ook over de werking van de technologie. Door hier al vroeg mee te beginnen zorg je ervoor dat de nieuwe generaties weerbaarheid ontwikkelen en daarmee grip behouden op het gebruik van generatieve AI in de (digitale) media.

Authenticiteit

Authenticiteit is echtheid en maakt iets daardoor betrouwbaar. Dit hoeft niet overeen te komen met de werkelijkheid en staat daarom niet gelijk aan de waarheid. De komst van Generatieve AI heeft ervoor gezorgd dat er maatschappelijke uitdagingen rondom authenticiteit zijn ontstaan. Binnen het veiligheidsdomein speelt bijvoorbeeld het probleem dat de *Chain of Evidence* niet te achterhalen is bij het gebruik van gesloten technologie. Dat maakt dat bewijsvoering lastig gemaakt wordt. Dit speelt ook in de media. Bij gebruik van AI voor het genereren van informatie wordt het lastig om als gebruiker te achterhalen wat de herkomst van de informatie is. Hetzelfde geldt voor de producent: als AI wordt ingezet als hulpmiddel, is het op het moment lastig in te schatten of hetgeen wat er gegenereerd wordt op waarheid berust. Dit maakt dat er vragen ontstaan rondom journalistieke authenticiteit. De vraag is hoe deze te borgen bij het gebruik van generatieve AI. Voorbeelden van mogelijke oplossingen zijn transparantie, *Explainable AI*, het borgen van publieke waarden in het ontwerp van de technologie of *human-in-the-loop AI*. Echter, het probleem is dat de *Big Tech* bedrijven en statelijke actoren niet de urgentie voelen om hierop in te zetten.

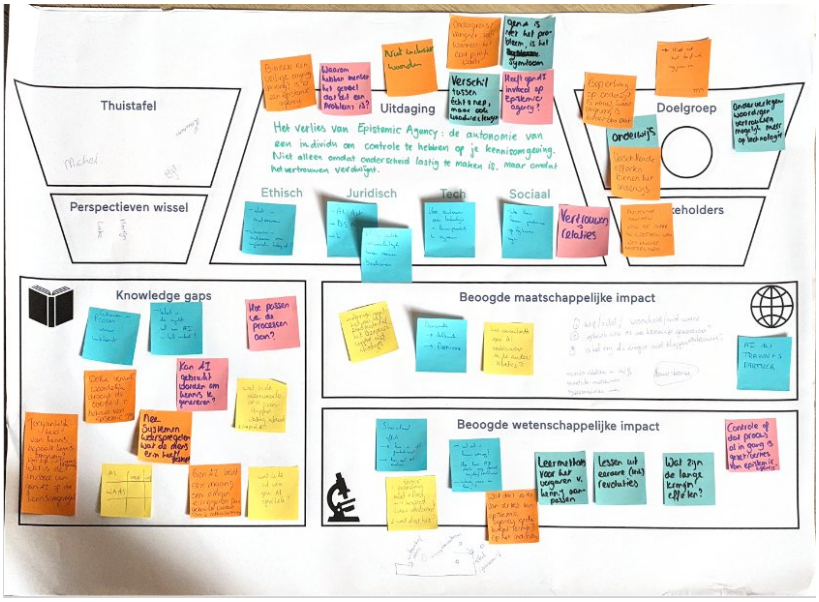
Na de analyse van de interviews en afstemming met het routemanagement, is er een co-creatie bijeenkomst georganiseerd waarin deze hoofdonderwerpen verder zijn uitgewerkt. De deelnemers van de bijeenkomsten waren ofwel de geïnterviewden of collega's van geïnterviewden die er tijdens de bijeenkomst niet zelf bij konden zijn. Daarnaast waren er nog enkele stakeholders aanwezig die we niet eerder in een interview hadden kunnen spreken. In totaal waren er 12 deelnemers.

Figuur 1 Co-creatie bijeenkomst

Aan de hand van een interactieve werkvorm zijn drie hoofdonderwerpen nader uitgewerkt: Epistemic Agency, Authenticiteit en Misbruik van Generatieve AI. De twee overige onderwerpen kwamen daarbinnen aan de hand van vragen op de gebruikte templates nog altijd aan bod. Per groepje werd onderling het voorliggende onderwerp uitgediept, de daarbij behorende kennisleemtes uitgewerkt en de beoogde maatschappelijke en wetenschappelijke impact opgehaald. Ook werd de deelnemers gevraagd welke doelgroep het onderzoek zich op zou moeten richten en welke stakeholders er bij het onderzoek betrokken zouden moeten worden. Er vond een perspectievenwissel plaats waarbij groepjes reflecteerden op elkaars input en later deze inzichten weer meenamen naar hun oorspronkelijke hoofdonderwerp ("thuistafel"). Afsluitend zijn de uitgewerkte thema's aan elkaar gepresenteerd en konden deelnemers een stem uitbrengen op het thema met de hoogste prioriteit.

De bijeenkomst leverde interessante gesprekken op, waarbij het ging over het borgen van authenticiteit in het ontwerp van generatieve AI, authenticiteit ≠ waarheid, de impact van generatieve AI op het individu en sociale relaties, *fact-checking*, effecten op de lange termijn bij gebrek aan authenticiteit en/of wantrouwen, kennisinfrastructuren, de impact op ons vermogen om kennis te beoordelen, wet- en regelgeving, de rol van de overheid en onderwijs. Door de verscheidenheid aan achtergrond van de deelnemers ontstonden er interessante, waardevolle en ook brede gesprekken die gedocumenteerd zijn aan de hand van templates en een prioritering. Op basis hiervan zijn we met het routemanagement in gesprek gegaan en heeft dit geleid tot een thema waar we in het

volgende hoofdstuk op ingaan.



Figuur 2 Voorbeeld van ingevulde template

Conclusie en aanbevelingen

Veel van de onderwerpen die in de interviews en de co-creatie sessie naar voren kwamen raken aan elkaar en dat maakt dat het uiteindelijke thema veel aspecten samenbrengt. Eigenlijk kwam alles samen binnen de focus op *Epistemic Agency* en de ontwikkeling daarvan ten aanzien van Generatieve AI in de digitale media. Het doel van onderzoek wat de NWA-ORC call mogelijk maakt, zou het stimuleren van digitale weerbaarheid tegen de ongewenste consequenties van de alomtegenwoordigheid van AI-gegenereerde content moeten zijn. Daarbij gaat het zowel over onbedoelde consequenties (zoals verwarring over wat waar is en wat niet) als misbruik van generatieve AI.

Binnen dit thema zou het onderzoek moeten verkennen hoe we als mensen *Epistemic Agency* kunnen ontwikkelen ten aanzien van generatieve AI in de digitale media. Dit houdt in dat ze de vrijheid en autonomie hebben (en voelen) om digitale content gegenereerd door AI te kunnen beoordelen op waarheid en de content daarmee vertrouwen. Het doel is het stimuleren van publieke weerbaarheid tegen mogelijk misbruik van generatieve AI. Mogelijk misbruik kunnen we niet voorkomen, de vraag is hoe we ermee omgaan. Het onderzoek op dit thema zou daarom moeten leiden tot

handelingsperspectief: wat kunnen mensen zelf doen om door AI gegenereerde content te kunnen beoordelen op waarheid. De oplossing daarvoor ligt niet in één domein. Het vraagt om een interdisciplinaire aanpak waarbij er zowel aandacht gegeven wordt aan de mens als de technologie. Daarbij ligt de focus op twee richtingen die elkaar zouden moeten aanvullen:

- **AI-geletterdheid**. Het verwerven van kennis en bewustzijn over de werking en gevolgen van generatieve AI. Het is nodig om meer in te zetten op kritische digitale vaardigheden en mediawijsheid. Daarbij is het niet alleen van belang hoe je door AI-gegenereerde content kunt herkennen, maar ook hoe de technologie werkt en hoe er mogelijk misbruik van gemaakt kan worden. Niet alleen scholen hebben de verantwoordelijkheid om AI-geletterdheid onder de bevolking te vergroten, maar ook binnen andere sectoren zou daar meer aandacht aan besteed moeten worden.
- **Technologie**. Het inzetten op tools die de herkomst van digitale content inzichtelijk maken of die gebruikers in staat stellen om de "waarheid" van AI-gegenereerde content beter in te schatten en te controleren. Detectie alleen is geen houdbare oplossing – er vindt een continue wapenwedloop plaats in de ontwikkeling van generatieve AI en technologie voor het kunnen herkennen daarvan. Echter, zonder detectie wordt het heel lastig om publieke weerbaarheid te bewerkstelligen. Transparantie over het gebruik van generatieve AI is daarmee ook niet voldoende. Het handelingsperspectief wat we mensen willen bieden gaat niet alleen om het kunnen herkennen van door AI gegenereerde content, maar ook over de mogelijkheden die men zelf heeft om de content te beoordelen op waarheid. Mogelijkheden hiervoor liggen bijvoorbeeld in *Explainable AI* of *Human-in-the-loop AI*. Voor een ontvanger moet het mogelijk zijn om te kunnen achterhalen waar de informatie vandaan komt, welk model er is gebruikt en op welke data het is getraind. Voor de producent van digitale content geldt hetzelfde. Als generatieve AI als hulpmiddel ingezet wordt, dan wil je kunnen achterhalen op het op waarheid gestoeld is.

Een aanbeveling is om het onderzoek specifiek op jongeren te richten. Tijdens zowel de interviews als de bijeenkomst is dit door veel stakeholders regelmatig genoemd als belangrijke doelgroep. Een leeftijdsgroep volop in ontwikkeling en met regelmaat geconfronteerd met digitale content. Door deze doelgroep als uitgangspunt te nemen voor het onderzoek, wordt er maatschappelijke meerwaarde gecreëerd met een zo lang mogelijke impact. Bovendien blijven de opgedane inzichten langer actueel. Het weerbaar

maken van een jonge generatie maakt dat de kans op mogelijk misbruik van generatieve AI in de toekomst (naar verwachting) steeds kleiner wordt.

In generatieve AI schuilt het risico dat de modellen gebaseerd zijn op over gerepresenteerde groepen, waardoor onevenwichtige structuren in de technologie in stand worden gehouden. Bovendien maakt de *black box van AI* dat deze onevenwichtige structuren slecht zichtbaar zijn. Het is daarom belangrijk om in het onderzoek naar technologie bewust rekening te houden met onder gerepresenteerde groepen. Wanneer zij profiteren van de kansen die de technologie biedt en beschermd worden voor de negatieve gevolgen. Door onder gerepresenteerde groepen als uitgangspunt te nemen, heeft het onderzoek en de inzichten die dat oplevert uiteindelijk maatschappelijke meerwaarde voor de hele samenleving.

Colofon

Uitgave
Waag Futurelab 15 oktober 205

Auteurs/redactie
Jikke van den Ende en Simone van der Burg

Contactgegevens
Waag Futurelab
Sint Antoniesbreestraat 69
1011 HB Amsterdam