Misbruik AI kan wetenschap zwaar beschadigen

DEEPFAKE SCIENCE – EEN DUISTER SCENARIO

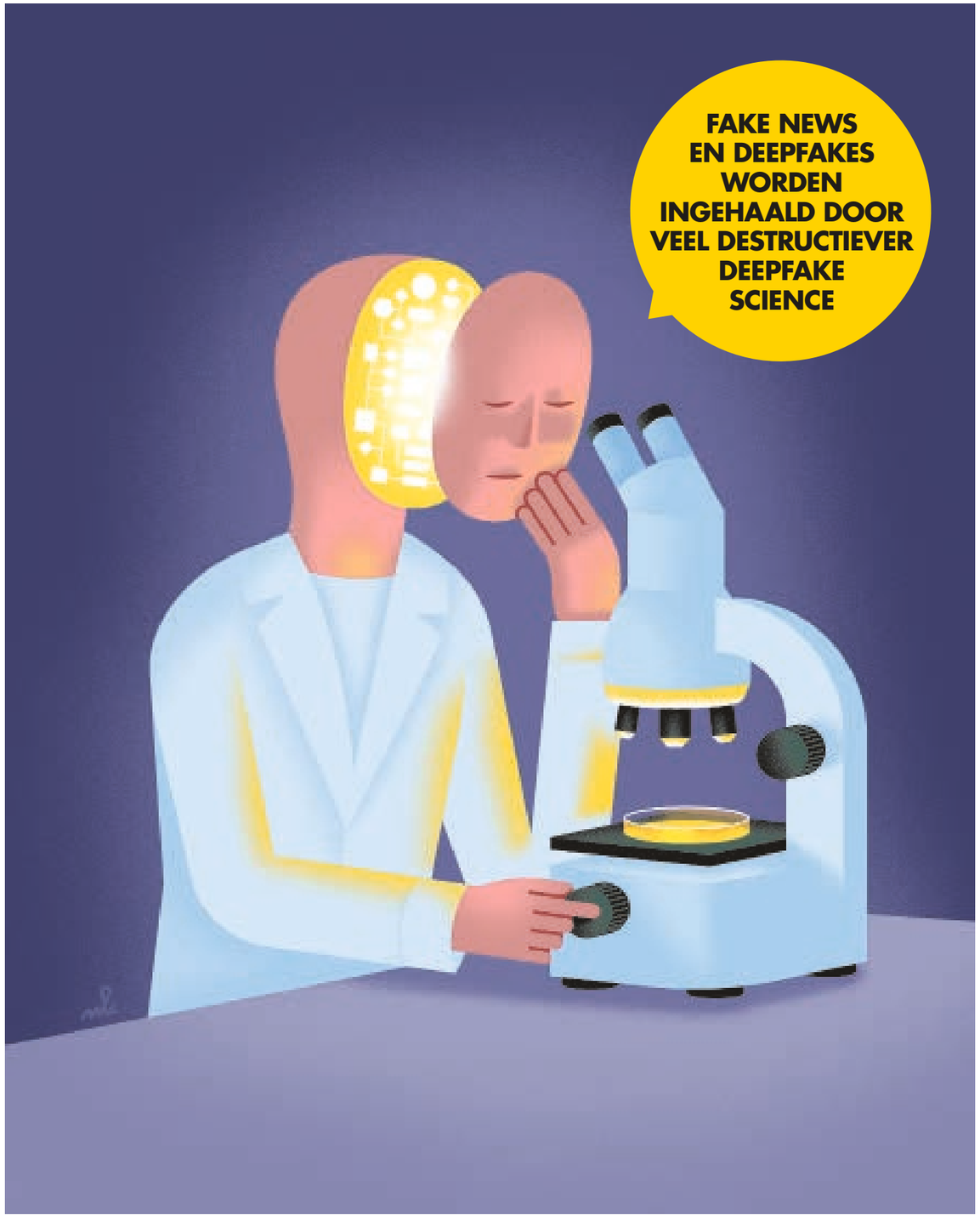
DAT AI KAN ZORGEN VOOR FAKE NEWS EN DEEPFAKES IS ALOM BEKEND. MOGELIJK NOG DESTRUCTIEVER IS DAT MALAFIDE GEBRUIK VAN AI OOK KAN BIJDRAGEN AAN DEEPFAKE SCIENCE. DAT KAN HET VERTROUWEN IN WETENSCHAP EN WETENSCHAPSGEBASEERD BELEID ZWAAR BESCHADIGEN EN ZEER DUISTERE GEVOLGEN HEBBEN VOOR GROTE MAATSCHAPPELIJKE VRAAGSTUKKEN. HET IS HOOG TIJD ONS DAAR GRONDIG TEGEN TE WAPENEN, CONSTATEREN LEON KESTER EN BART WERNAART.

door Leon Kester en Bart Wernaart beeld Marc Kolle

TEN MINSTE DRIE BELANGRIJKE ONTWIKKELINGEN HEBBEN MAATSCHAPPELIJKE IMPACT EN ROEPEN DUS VRAGEN OP: Artificial Intelligence (AI); Extended Reality (XR) en Epistemische Veiligheid (EV), een begrip van Alan Turing dat de veiligheid van kennis aanduidt. Deze gebieden werden tot voor kort vaak los van elkaar onderzocht in een ethische context. De laatste jaren dringt het besef mondjesmaat door dat juist de combinatie van deze technologische ontwikkelingen voor veel grotere maatschappelijke risico's kan zorgen. Platgeslagen: wanneer technologie 'intelligente' systemen behelst (AI), die bepalend zijn in hoe mensen de wereld zien, ervaren (XR) en bepalend zijn in hoe zeker zij zich daarbij tegen epistemische bedreigingen voelen (EV) dan moeten we ons ernstige zorgen maken en acuut handelen, alle mogelijkheden en techno-optimisme ten spijt. Een voorbeeld waar deze drie ontwikkelingen samen komen is deepfake science. Waar we het vooralsnog veel hebben over fake news, deepfakes en het herkennen van nep-berichten en visuals, halen de technologische mogelijkheden dergelijke discussies snel in. Deepfake science is een veel destructiever mogelijkheid.

ONDERUIT HALEN

AI wordt al ingezet om teksten te schrijven, denk bijvoorbeeld aan NBD Biblion die onlangs al haar recensenten verving door AI of de toenemende praktijk van AI-ondersteunde journalistiek. Dit roept vragen op rondom AI-veiligheid, intellectueel eigendom, autonomie, en gerelateerde (economische) belangen. Maar



Als verdediging moeten we technologie gebruiken die onze creativiteit stimuleert

deze ontwikkeling staat in schril contrast met de potentie van AI als ondersteuner of zelfs als auteur in de wetenschap. Waar we in theorie fake news nog langs de meetlat van de wetenschap kunnen leggen, wordt het lastig wanneer AI moedwillig gebruikt wordt als een politiek gedreven wapen om wetenschappelijke standpunten onderuit te halen. Mogelijk malafide gebruik kan bijvoorbeeld bestaan uit het kunstmatig creëren van datasets, video's en audio's of geschreven wetenschappelijke artikelen. De taalkundige imitatiecapaciteit van AI is drastisch verbeterd met bijvoorbeeld GPT-3-achtige systemen. In theorie zou het binnen niet afzienbare tijd mogelijk zijn epistemische kwetsbaarheden in de wetenschap zo te misbruiken.

POLARISATIE

In de digitale infrastructuur waarin ook wetenschappelijk werk wordt gedeeld, spelen socialemediaplatforms een belangrijke rol. Deze zijn in handen van een beperkt aantal private spelers, die financieel doorgaans wel varen bij polarisatie. De plannen voor een metaverse leiden tot een steeds nadrukkelijkere integratie van XR in ons dagelijkse leven. Dat heeft niet alleen effect op hoe we de wereld om ons heen ervaren en interacteren met anderen, maar ook op hoe we informatie zien en beleven. Het is mogelijk dat een toekomstige aanval met deepfake science doorsijpelt in onze metaverse-beleving. Het wordt daarbij in toenemende mate lastiger om het individu te beschermen tegen dergelijke aanvallen. Waar wetenschappelijk bewijs soms een speelbal lijkt van politieke stromen, wordt dit door deze technologische ontwikkelingen een veel groter probleem, waarbij het vertrouwen in wetenschap en overheidsbeleid gebaseerd op wetenschap zwaar beschadigd zou kunnen worden. Dit kan duistere gevolgen hebben voor grote vraagstukken, zoals klimaatverandering, epidemieën, migratie of de werking van democratie. Tenzij we ons wapenen tegen deze toekomstige praktijk.

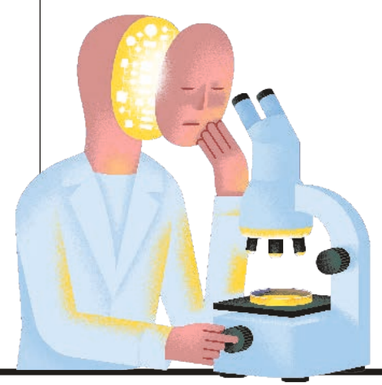
BETERE VRAGEN STELLEN

Wat kan technologie (met name AI) wel en wat niet? Hoe verhoudt zich dat tot (het doel van) wetenschap? Laten we niet naïef zijn: uiteindelijk kan alles wat imiteerbaar is, geïmiteerd worden door technologie. Het enige wat technologie niet oppervlakkig kan imiteren is de unieke capaciteit van mensen om nieuwe – nog niet bestaande - theorieën te genereren. Daarin onderscheidt uiteindelijk een wetenschapper zich van wetenschapstechnologie. Mensen zijn lang niet altijd in staat om authentiek van nep te onderscheiden wanneer men geconfronteerd wordt met AI-gegenereerde teksten, menselijke teksten, of een blend van beiden. Bovendien hoeven AI-gegenereerde teksten niet per se slechte ideeën in mensen op te wekken, alleen doordat deze niet door mensenhand geschreven werden.

Dat het doel van wetenschap is om de waarheid te achterhalen, wordt al geruime tijd door ethici en wetenschapsfilosofen verworpen. Mensen kunnen hooguit een perceptie van wat ze achterhalen tot zich nemen. Het doel van wetenschap is het steeds beter verklaren van datgene wat we onderzoeken. De vraag is dus niet of wetenschap door mensen of machines gegenereerd wordt (de bron), maar of wetenschap leidt tot betere of slechtere verklaringen (de inhoud) van datgene wat we onderzoeken. Omdat het als mens steeds lastiger wordt echt van nep te onderscheiden, ligt het voor de hand ons heil te zoeken in bijvoorbeeld het trainen van AI hiervoor. Maar het enige wat hedendaagse AI kan, is imiteren. De zonderling zal hierdoor vrij eenvoudig het label 'nep' kunnen krijgen. Kortom, als wij erop focusseren echt van nep te onderscheiden, bestaat het risico dat de nieuwe Einsteins of Hawkins van deze wereld geen wetenschappelijk podium meer krijgen. Als een 'bad actor' deepfake science wil genereren zal het meestal resulteren in een minder goede verklaring in vergelijking met de beste al bekende verklaring. Samengevat: voor een robuustere verdediging tegen deepfake science attacks moeten we ons richten op het onderscheiden van 'beter' van 'slechter' (in plaats van 'echt' van 'nep') en onze technologie gebruiken om onze eigen creativiteit te stimuleren.

MODERNE GRONDRECHTENDIALOOG

Veel grondrechten (denk aan privacy, vrijheid van meningsuiting, het recht op informatie of intellectueel eigendom) zijn vastgesteld voor een wereld waarin men zich grotendeels offline bewoog. Deepfake science attacks staan niet per se in de weg van het primaire doel van een groot tech-bedrijf: winstoptimalisatie. Als grote techbedrijven een verdienmodel kunnen maken van polarisatie onder het motto 'vrijheid van meningsuiting' en geen inzage willen geven in de manier waarop hun platforms technologisch werken onder het mom van 'intellectueel eigendom', dan vormt dit een toepassing van grondrechten die niet in het algemeen belang is. Het wordt dat voor de gemiddelde burger zeer lastig om op een daadwerkelijk autonome manier invulling te geven aan datzelfde recht om een mening te uiten, maar ook aan het recht op privacy of informatie.



URGENTIE

Daarbij moeten we goed nadenken over democratische processen in de snelle technologische ontwikkelingen. Hierbij is een belangrijke rol weggelegd voor wetgevers (top-down), die zich vooralsnog weinig bekwaam voelen zich op dit terrein te begeven. Het wordt tijd dat wetgevers op alle niveaus, maar zeker op landelijk en Europees niveau, de urgentie inzien van het strikter reguleren van technologie, waarbij veel aandacht zou moeten zijn voor niet onderhandelbare veiligheidscriteria. In diezelfde wetgeving zou ook aandacht moeten zijn voor de inbedding van moraldesign-processen (bottom-up), waarbij menselijke waarden op een systematische manier vertaald worden naar designprincipes van nieuwe technologische toepassingen. Wetgeving is altijd trager dan innovatie, het is dus noodzakelijk om juist in de ontwerpfase van nieuwe technologie de stem van de samenleving al te verankeren door burgers te laten experimenteren met prototypes. Ook als het ons voorstellingsvermogen nu misschien te buiten gaat: als deepfake science een serieuze bedreiging kan zijn, hoe zou het dan zijn met deepfake religion, deepfake elections of deepfake education? Tenslotte, zeker gezien de op handen zijnde ontwikkelingen in kwantumtechnologie ligt het voor de hand een versnelling van dit soort vraagstukken te verwachten. Laten we ons dus goed voorbereiden, kwalitatief betere discussies voeren en op een moderne manier naar grondrechten kijken.

De auteurs spreken hun dank uit aan Nadisha-Marie Aliman, postdoctoraal gastonderzoeker bij de Universiteit Utrecht, voor haar feedback en kritische blik bij het tot stand komen van dit stuk.

REACTIES EN BIJDRAGEN

Voor reacties en nieuwe bijdragen van IT-experts:
Tanja de Vrede
020-2467230
t.d.vrede@agconnect.nl

AUTEUR

LEON KESTER
is Senior Research Scientist bij TNO op het gebied van AI(XR) Strategy, AI(XR) Safety, AI(XR) Meta-ethics, AI(XR) Governance.

BART WERNAART
is lector Moral Design Strategy bij Fontys Hogescholen.