AI-criminaliteit: een verkenning van actuele verschijningsvormen

Marc Schuilenburg en Melvin Soudijn*

De lancering van de veelbesproken chatbots ChatGPT en Bard Gemini heeft niet alleen de interesse van legale partijen gewekt, maar ook die van criminelen. Zo berichtte BBC News dat op basis van ChatGPT er chatbots in omloop zijn die e-mails, teksten en andere berichten op sociale media kunnen genereren en versturen om personen op te lichten.[1] Zulke bots staan bekend onder veelzeggende namen als Crafty Emails, WormGPT en FraudGPT. Canadese onderzoekers ontdekten ook het bestaan van DarkBART, een door criminelen zelfgemaakte darknetversie van Googles Bard Gemini, waarmee phishingcampagnes kunnen worden opgezet en malware kan worden gemaakt en verspreid.[2] OpenAI, het Amerikaanse bedrijf achter ChatGPT, kwam met de volgende reactie op de onthullingen: 'We don't want our tools to be used for malicious purposes, and we are investigating how we can make our systems more robust against this type of abuse.'[3] Inmiddels hebben AI-toepassingen hun weg gevonden naar alle domeinen in de samenleving. AI wordt niet alleen in de private sector gebruikt, ook publieke partijen zetten AI in om hun dienstverlening te verbeteren of om bepaalde taken efficiënter te kunnen uitoefenen. Zo gebruikt de Nationale Politie verschillende AI-tools in het kader van de opsporing en handhaving van criminaliteit, waaronder toepassingen die zowel 'voorspellend' als 'in real time' en 'retrospectief' kunnen worden ingezet (Schuilenburg & Soudijn 2023). Zowel private als publieke partijen maken daarbij gebruik van verschillende typen algo-

* Prof. dr. mr. M.B. Schuilenburg is hoogleraar Digital Surveillance aan de Erasmus Universiteit Rotterdam. Dr. M.R.J. Soudijn is senior onderzoeker bij de Eenheid Landelijke Opsporing van de Nationale Politie. De bijdrage van Melvin Soudijn is op persoonlijke titel geschreven. Beide auteurs danken Thijs van de Bult en Jaap de Waard voor het verzamelen van relevante literatuur en hun waardevolle suggesties.

1 Zie https://securityboulevard.com/2023/08/after-wormgpt-and-fraudgpt-darkbert-and-darkbart-are-on-the-horizon/, bezocht op 19 februari 2024.

2 Zie https://slashnext.com/blog/ai-based-cybercrime-tools-wormgpt-and-fraudgpt-could-be-the-tip-of-the-iceberg/, bezocht op 19 februari 2024.

3 Zie www.bbc.com/news/technology-67614065, bezocht op 19 februari 2024.

doi: 10.5553/JV/016758502024050001002

ritmen, van eenvoudige toepassingen, waaronder beslisbomen en data-uitwisselingssystemen, tot technisch zeer complexe toepassingen, zoals *machine learning*, waarbij het algoritme zijn eigen weg gaat en zich kan onttrekken aan het zicht en vermogen van betrokken professionals. Hoewel de intelligentie van AI een ingewikkeld onderwerp is waarover veel wordt gediscussieerd, gaat het bij AI vooral om deze zeer complexe toepassingen, in het bijzonder om systemen die – met een zekere mate van zelfstandigheid – acties ondernemen om bepaalde doelen te bereiken.[4]

Hoewel de maatschappelijke discussie over AI en de rol van algoritmen in onze samenleving is losgebarsten, is dat nog amper zichtbaar op de criminologische radar. AI in relatie tot het *plegen* van criminaliteit krijgt weinig tot geen aandacht. Dit terwijl de doorwerking en inzet van AI het speelveld van cybercriminaliteit aanzienlijk kunnen verruimen. AI-toepassingen zoals gratis toegankelijke chatbots hebben namelijk de drempel verlaagd voor personen met criminele intenties. Dat komt omdat voor het gebruik van deze chatbots geen hoog kennisniveau nodig is. Daarmee kan in potentie zowel het dader- als het slachtofferschap toenemen. Ook kan AI leiden tot nieuwe vormen van criminaliteit, die maatschappelijk meer schade kunnen aanrichten dan al langer bestaande vormen van cybercriminaliteit (denk aan online fraude en cyberpesten). Maar met welke vormen van AI-criminaliteit moet er rekening worden gehouden?

In dit artikel geven we op basis van een uitgebreide literatuurstudie een overzicht van de meest actuele verschijningsvormen van AI-criminaliteit. Hiertoe onderscheiden we drie vormen van AI-criminaliteit: (1) met AI, (2) gericht op AI en (3) door AI. Voordat we ingaan op deze indeling bespreken we eerst hoe AI het veiligheidsdomein verandert. Vervolgens gaan we in op de resultaten van het literatuuronderzoek naar AI-criminaliteit. In de conclusie duiden we de verschillen tussen AI-criminaliteit en cybercrime en bespreken we tot welke vragen en uitdagingen AI-criminaliteit leidt voor de preventie en opsporing ervan.

4 Het is zeer lastig om een eenduidige beschrijving te geven van AI. Wij houden in dit artikel de definitie van de High-Level Expert Group on Artificial Intelligence van de Europese Commissie aan: 'Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals' (2019).

doi: 10.5553/JV/016758502024050001002

AI in relatie tot criminaliteit

We leven in een digitale tijd waarin steeds meer aspecten van ons leven zich online afspelen en het onderscheid tussen fictie en werkelijkheid steeds lastiger is te maken. Digitalisering laat zich hierin makkelijk verbinden met de meest recente ontwikkelingen op het gebied van criminaliteit. Simpel gezegd, de samenleving digitaliseert en daarmee ook de criminaliteit. In de literatuur wordt hiertoe een onderscheid gemaakt tussen cybercriminaliteit in enge zin en computergerelateerde delicten. Cybercriminaliteit in enge zin duidt op het plegen van strafbare feiten waarbij netwerken en computers en de daarin opgeslagen gegevens zowel het doelwit als het middel zijn van criminaliteit, zoals in het geval van hacken ('computervredebreuk'), gijzelsoftware en het verspreiden van kwaadaardige virussen ('malware'). Computergerelateerde (ook wel genoemd: gedigitaliseerde) criminaliteit gaat om klassieke misdrijven die (mede) worden gepleegd met behulp van computers. Je kunt hierbij denken aan misdrijven als witwassen, zedendelicten en drugshandel die worden gepleegd via internet, e-mail of app (o.a. Oerlemans & Van der Wagen 2021; Schermer 2022).

Uit de meest recente cijfers blijkt dat in westerse landen de geregistreerde criminaliteit sinds 2002 met ruim een kwart is gedaald. Maar die daling gaat niet op voor cyber- en gedigitaliseerde criminaliteit. Het slachtofferschap hiervan neemt flink toe en de impact op de samenleving is groot. Zo is in 2022 15% van de Nederlanders van 15 jaar of ouder het slachtoffer geweest van een of meer vormen van online criminaliteit. Dat zijn ruim 2 miljoen personen (CBS 2023). Cijfers van het CBS tonen aan dat het vooral gaat om oplichting en fraude, gevolgd door hacken en bedreiging en intimidatie. Naar verwachting zal de komende jaren de omvang van cyber- en gedigitaliseerde criminaliteit verder toenemen, waarbij Nederlanders – als het gaat om cyberveiligheid – zich vooral zorgen maken om het misbruik van hun bank- en persoonsgegevens.

Waar AI aanvankelijk nog als onderdeel van de bredere trend van digitalisering werd gezien, komt er steeds meer aandacht voor de centrale rol van AI in de samenleving en de transformerende uitwerking ervan voor hoe we leven, wonen en werken. Algoritmen stellen diagnoses in de zorg, voorspellen fraude met sociale voorzieningen, besturen zelfrijdende auto's, handelen op de beurs ('algotrading'), maken digitale

doi: 10.5553/JV/016758502024050001002

vonnissen, doen echtheidsonderzoeken aan schilderijen, sturen onze persoonlijke voorkeuren op sociale media en herkennen verdachte personen in de wijk (Ring videodeurbel van Amazon). AI zit met andere woorden in de haarvaten van de samenleving en vormt de motor van ons bestaan. De Wetenschappelijke Raad voor het Regeringsbeleid (WRR 2023) spreekt daarom van een 'systeemtechnologie', een uitvinding met een systematisch effect door heel de samenleving heen, zoals elektriciteit dat was in de negentiende eeuw en de verbrandingsmotor in de vorige eeuw. 'Systeem' heeft hierin een dubbele betekenis. Enerzijds maakt AI onderdeel uit van een breder systeem van data en hardware. Anderzijds, zo schrijft de WRR, heeft 'AI een effect op allerlei systemen en processen in onze samenleving' (2023, p. 128).

Kijken we naar de veiligheidspraktijk, dan prevaleert een technisch-economische benadering als het gaat om het gebruik van AI en algoritmen. AI wordt vooral gezien als een manier om een belangrijke efficiëntieslag te maken in de preventie en opsporing van criminaliteit. Het gaat daarbij om zaken zoals de snelheid waarmee zeer grote hoeveelheden data kunnen worden verzameld en geanalyseerd, of efficiencyargumenten dat door het gebruik van AI-toepassingen zowel de werkprocessen als de opsporing van criminaliteit grondig verbeteren (Schuilenburg 2024). Maar hoewel de opsporing van criminaliteit profijt kan hebben van AI-toepassingen, brengt AI ook nieuwe risico's met zich mee voor de veiligheid. Een belangrijke vraag daarbij is wat de komst van AI betekent voor de aard en omvang van criminaliteit. AI-toepassingen kunnen namelijk ook voor criminele doeleinden worden ingezet. Anders gezegd, AI kan door criminelen worden gebruikt en worden gezien als een volgende stap in de digitalisering van criminaliteit. In de volgende paragrafen gaan wij hierop dieper in.

AI-criminaliteit: drie verschijningsvormen

Op basis van een literatuurstudie naar de relatie tussen criminaliteit en AI onderscheiden wij drie verschijningsvormen: (1) met AI, (2) gericht op AI en (3) door AI.[5] Daarbij blijkt dat er soms sprake is van

5 Zie ook King e.a. (2020), Hayward en Maas (2020) en het rapport *AI-enabled future crime*, zie www.ucl.ac.uk/steapp/collaborate/policy-impact-unit/policy-brief-ai-enabled-future-crime, bezocht op 19 februari 2024.

doi: 10.5553/JV/016758502024050001002

enige overlap tussen deze vormen. Het is niet onze bedoeling om alle verschijningsvormen uitputtend te behandelen, maar om een eerste indruk te geven van vormen van AI-criminaliteit die er momenteel al zijn of waarvan verwacht wordt dat zij in de loop van de tijd een hoge vlucht zullen nemen.

Criminaliteit met AI

In deze categorie gaat het om verschijningsvormen waarbij AI als hulpmiddel wordt gebruikt om traditionele vormen van criminaliteit te plegen. Zonder AI zouden deze vormen van criminaliteit dus ook blijven bestaan.
In de literatuur lijken vooral *deepfakes* (een samentrekking van de woorden *deep learning* en *fakes*) zorgen te baren en als de grootste dreiging voor de veiligheid te worden gezien. Deepfaketechnologie biedt de mogelijkheid om bestaande afbeeldingen en bewegende beelden te combineren en over elkaar heen te zetten. Met behulp van AI kunnen zo compleet nieuwe beelden worden gegenereerd die met het blote oog niet of nauwelijks van echt zijn te onderscheiden. Hierdoor lijkt het alsof iemand iets zegt of doet, wat in werkelijkheid nooit is gebeurd (o.a. Maras & Alexandrou 2018; Custers 2021; Blauth e.a. 2022; Mai e.a. 2023). Hoewel de manipulatie van beelden niet nieuw is, denk aan programma's zoals Photoshop, worden door AI de mogelijkheden om beeldmateriaal te manipuleren steeds beter, gebruiksvriendelijker en toegankelijker.
Deepfakes kunnen worden gebruikt voor strafbare delicten als kinder- en wraakporno, voyeurisme, misleiding, oplichting en fraude. Een voorbeeld hiervan zijn bekende personen die in deepfakefilmpjes je aansporen om geld te investeren in frauduleuze projecten. Een ander voorbeeld is het vervaardigen, toegankelijk maken en verspreiden van gemanipuleerde naakt- of seksbeelden (*deepnudes*) door bijvoorbeeld het gezicht van een persoon te projecteren op pornografische beelden van een ander persoon met als doel iemands reputatie te besmeuren, personen af te persen, of iemand angst aan te jagen. Prominente politici, muzikanten en acteurs die hiervan het slachtoffers zijn geworden, zijn onder meer Meghan Markle, Michelle Obama, Natalie Portman,

doi: 10.5553/JV/016758502024050001002

Taylor Swift en Scarlett Johansson.[6] In Nederland is presentatrice en journaliste Welmoed Sijtsma slachtoffer geworden van zo'n deepfake-pornovideo. De maker hiervan is door de rechtbank tot een taakstraf van 180 uur veroordeeld.[7]

Ook wordt kunstmatig gegenereerde kinderporno gefabriceerd. Hoewel de beelden niet altijd échte kinderen hoeven te bevatten, kunnen deze deepfakes schadelijk zijn omdat zo de cognitieve associatie tussen seksualiteit en kinderen kan worden versterkt (Ratner 2021; Faassen e.a. 2021). In de literatuur wordt daarnaast gewezen op het risico van gemanipuleerd beeld- en geluidmateriaal dat als overtuigend bewijs wordt gepresenteerd in de strafrechtspraktijk, met als mogelijk gevolg een afnemend vertrouwen in de rechtspraak (Maras & Alexandrou 2018). Ook zijn er speech-deepfakes waarbij beelden worden gecombineerd met kunstmatige stemmen die zijn gegenereerd via AI om onder meer personen op te lichten (Mai e.a. 2023).

Voice cloning is een andere verschijningsvorm van traditionele criminaliteit die wordt gepleegd met AI. Hierbij wordt gedaan alsof iemand iets heeft gezegd, terwijl dat niet zo is. Zo kan de stem van een familielid via AI worden nagebootst en worden gebruikt om personen telefonisch op te lichten (Jeong 2020). Niet alleen kunnen familieleden worden opgelicht, bijvoorbeeld door een verhaal te vertellen waarin iemands kind in nood verkeert en onmiddellijk geld nodig heeft, ook bekende personen, onder wie leidinggevenden van bedrijven, zijn doelwit geworden van stemdiefstal vanwege hun openbare profiel op sociale media. Zo hoorde de Britse acteur en schrijver Stephen Frye zichzelf als voice-over van een documentaire waarvan hij niets afwist. In een reactie stelde Frye: 'It could have me read anything from a call to storm parliament to hard porn, all without my knowledge and without my permission.'[8]

Een combinatie van deepfakes, *voice cloning* en door AI gecreëerd nepnieuws kan er ook toe leiden dat er bewust sprake is van manipulatie van de publieke opinie, het zaaien van verwarring of het ondermijnen van het vertrouwen in instituties. Zo wordt verwacht dat de

6 Neurowetenschapper Erik Hoel voorspelt dat elke vrouw die bekend is bij een breder publiek, er rekening mee moet houden dat waarschijnlijk deepfake porno van haar wordt gemaakt. https://www.theintrinsicperspective.com/p/here-lies-the-internet-murdered-b, bezocht op 11 maart 2024.

7 Rb. Amsterdam 2 november 2023, ECLI:NL:RBAMS:2023:6923.

8 Zie www.theguardian.com/technology/2023/sep/20/it-could-have-me-read-porn-stephen-fry-shocked-by-ai-cloning-of-his-voice-in-documentary, bezocht op 19 februari 2024.

doi: 10.5553/JV/016758502024050001002

verspreiding van desinformatie door de razendsnelle opkomst van AI een van de belangrijkste dreigingen wordt in zowel conflictsituaties als de komende verkiezingsjaren, waarin de helft van de wereldbevolking kan stemmen.[9] Daarbij moet in het achterhoofd worden gehouden dat het door AI goedkoper en makkelijker dan ooit is geworden om met misleidende informatie de tegenpartij in een kwaad daglicht te stellen. In dat kader wordt gesproken van *botshit* – bullshit die door bots wordt gegeneerd (Hannigan e.a. 2023).[10]

Buiten de militaire en politieke arena's kan *botshit* worden ingezet voor *pump and dump schemes* en andere vormen van frauduleuze activiteiten om financiële markten te manipuleren. Zo prijzen chatbots op sociale media bepaalde aandelen, die niet op de gereguleerde marktindex genoteerd staan, aan via misleidende nieuwsverhalen of zogenaamde aanbevelingen van bekende personen. Vervolgens stappen nietsvermoedende particulieren in, waardoor de koers stijgt. Direct na deze koersstijging verkopen ('dumpen') de fraudeurs de aandelen en blijft de nietsvermoedende koper met waardeloze aandelen zitten en is zijn of haar geld kwijt.

Bij *phishing* gaat het om *Large Language Models* (LLM) als ChatGPT die worden gebruikt om onder meer (gepersonaliseerde) nepmails of websites te fabriceren om personen om de tuin te leiden. Europol (2023) schrijft in haar rapport *ChatGPT. The impact of Large Language Models on law enforcement* dat dergelijke chatbots van onschatbare waarde zijn voor criminelen met weinig technische kennis die over malware willen beschikken. Deze chatbots schrijven zowel teksten voor oplichtingsmails als de codes voor websites waarmee gegevensdiefstal kan worden gepleegd.[11] Het Britse National Cyber Security Centre (2024) stelt in zijn nieuwste rapport: 'To 2025, GenAI and Large Language Models will make it difficult for everyone, regardless of their level of cyber security understanding, to assess whether an email or password reset request is genuine, or to identify phishing, spoofing or social engineering attempts.' Door AI gegenereerde *bad bots* kunnen

9 Zie www.nrc.nl/nieuws/2024/01/10/misleidende-botshit-vraagt-om-alertheid-in-het-cruciale-verkiezingsjaar-2024-a4186531, bezocht op 19 februari 2024.

10 Zie www.theguardian.com/commentisfree/2024/jan/03/botshit-generative-ai-imminent-threat-democracy, bezocht op 19 februari 2024.

11 Zie www.hyas.com/blog/blackmamba-using-ai-to-generate-polymorphic-malware, bezocht op 19 februari 2024. Een potentieel extra risico dat hierbij wordt genoemd, is dat malware die louter en alleen door AI is geschreven een ander soort logica volgt of onvoorspelbaarder is dan door mensen geschreven programma's. Het zou daardoor moeilijker door bestaande antivirusprogramma's worden herkend.

doi: 10.5553/JV/016758502024050001002

onder meer worden gebruikt om nepmails ongericht en meerdere keren naar grote groepen personen te sturen of om websites te overspoelen met meer berichten dan ze aankunnen (o.a. Boshmaf e.a. 2013; Seymour & Tully 2016).
Tot slot kan in de categorie 'Criminaliteit met AI' worden gedacht aan *AI assisted stalking*.[12] In zulke gevallen wordt AI ingezet als een surveillancetool. Een stalker zou met behulp van AI het gedrag van zijn of haar slachtoffer kunnen analyseren en voorspellen waar deze fysiek aanwezig zal zijn door de data die het slachtoffer op bijvoorbeeld sociale media achterlaat. Een stalker zou ook *voice cloning* kunnen gebruiken om dicht bij het slachtoffer te komen zonder dat deze het door heeft.

Criminaliteit gericht op AI

In de tweede categorie gaat het om criminaliteit die specifiek is gericht op AI-systemen. Het handelt met andere woorden om strafbare delicten die tegen AI-systemen worden gepleegd. Zo wordt de mogelijkheid genoemd om AI-systemen te hacken die autonoom tot beslissingen komen. Wij geven twee voorbeelden hiervan.
Het eerste voorbeeld betreft zelfrijdende auto's, die met AI en zonder chauffeur passagiers van A naar B brengen. Naast San Francisco wordt ook in andere steden geëxperimenteerd met dergelijke robottaxi's die volgepakt zijn met sensoren om de omgeving te verkennen. Zelfrijdende auto's verschaffen terroristen in potentie een nieuw wapen. Zo zou het centrale AI-systeem kunnen worden gehackt, om vervolgens de wagens zo te programmeren dat er terroristische aanslagen mee worden gepleegd (Caldwell e.a. 2020).
Een tweede voorbeeld is dat AI-systemen in de financiële sector doelwit kunnen zijn van criminelen (Blauth e.a. 2022). Ruim 80% van de handel op de Amerikaanse aandelenmarkt wordt inmiddels uitgevoerd door softwaretools die op algoritmen zijn gebaseerd.[13] De algoritmen voeren hierbij zelfstandig allerlei transacties uit op basis van bepaalde parameters en het verloop van de koersen. Als echter op grote schaal misleidende informatie wordt opgevoerd, zal dit de geautomatiseerde beslissingen beïnvloeden. Er kan dan een *flash crash* op de digitale

12 Zie www.cagoldberglaw.com/ai/, bezocht op 19 februari 2024.
13 Zie https://admiralmarkets.com/nl/educatie/artikelen/automatisch-forex-handelen/algorithmic-trading-strategies, bezocht op 19 februari 2024.

doi: 10.5553/JV/016758502024050001002

markt ontstaan, een plotselinge scherpe daling (gevolgd door een snel herstel) van de aandelenkoersen als gevolg van geautomatiseerde transacties die op elkaar reageren. Als de algoritmen bewust worden gemanipuleerd, wordt ook wel gesproken van *poisoning attacks*. Het systeem is dan gevoed met 'vuile data': data die óf onrechtmatig zijn verkregen óf onjuist zijn (Richardson e.a. 2019; Das & Schuilenburg 2020).

Criminaliteit door AI

Met deze categorie doelen wij op criminaliteit die zelfstandig door AI wordt gepleegd.[14] Het menselijk handelen is hierbij naar de achtergrond verdrongen. AI neemt beslissingen die naar de letter van de wet strafbaar zijn, wat de vraag oproept in welke mate AI kan worden gezien als een 'actor' in de uitvoering van criminaliteit. Kenmerkend voor deze verschijningsvorm is dat de criminaliteit vaak pas na enige tijd zichtbaar wordt, omdat er sprake is van een stil en sluipend proces. Dergelijke door AI gepleegde criminaliteit kan plaatsvinden in relatie tot de overheid, het bedrijfsleven of publiek-private samenwerkingen. We noemen hieronder enkele voorbeelden hiervan.
Als AI zelfstandig beslissingen neemt, bestaat er het risico op discriminatoir gedrag.[15] Het gaat hierbij om beslissingen waarbij het systeem onterecht onderscheid maakt tussen bepaalde bevolkingsgroepen bijvoorbeeld. De oorzaak is vaak simpel: de data waarop het systeem beslissingen maakt, zijn niet objectief, maar bevooroordeeld ('bias')

14 In de literatuur wordt hierbij doorgaans aan oorlogssituaties gedacht. AI kan bijvoorbeeld worden gebruikt om zwermen drones aan te sturen om doelen te bestoken of luchtverdediging te overweldigen. Een ander voorbeeld zijn de zogeheten 'killer robots', die in oorlogsgebieden autonome beslissingen nemen en zelfstandig doelen selecteren. In beide gevallen is er door deze *lethal autonomous weapon systems* (LAWS) risico op burgerslachtoffers (Brundage e.a. 2018; Horowitz 2019; Scharre 2015, 2016). Dit alles ligt echter meer in de lijn van het oorlogsrecht dan het strafrecht. Deze vormen van AI worden in dit artikel daarom verder buiten beschouwing gelaten.

15 Er bestaat ook risico op discriminatoir gedrag als de AI bewust wordt gestuurd. Zo bleek Google's Bard Gemini een overcorrectie op diversiteit toe te passen door *prompt engineering*. Zonder dat gebruikers er weet van hadden, hadden programmeurs van Gemini bij zoekopdrachten op de achtergrond een beslisregel toegevoegd om diversiteit uit te dragen. Het gevolg was dat de AI historische flaters sloeg door onder meer alleen Afrikaans uitziende Vikingen of een vrouwelijke paus te produceren. Omdat *prompt engineering* doelbewust ingrijpen vereist, worden zulke zaken verder niet in deze categorie opgenomen.

doi: 10.5553/JV/016758502024050001002

ten nadele van een bevolkingsgroep.[16] Ook leiden dergelijke modellen ertoe dat zelfbevestigende effecten optreden, waardoor een cirkelredenering ontstaat. De vuile data resulteren dan in maatregelen die negatief uitpakken voor een bepaalde groep. Die maatregelen leiden tot nieuwe vuile data, waardoor het negatieve aspect steeds opnieuw wordt benadrukt. Bij onvoldoende waarborgen leiden actuariële modellen zoals *predictive policing* (bijvoorbeeld 'crime mapping') en *predictive justice* tot grote risico's (Peeters & Schuilenburg 2018).[17] De burger kan zich hier slecht tegen verweren omdat deze geen inzage krijgt in het algoritme, vooral als deze modellen door commerciële partijen zijn ontwikkeld en tot bedrijfsgeheim worden bestempeld (Zuboff 2019; Slobogin & Brayne 2022). Maar ook eindgebruikers, zoals de overheid of de rechterlijke macht (en zelfs de eigen ontwikkelaars), hebben vaak onvoldoende tot geen besef van de precieze werking onder de motorkap door de iteratieve aanpassingen van het model (Das & Schuilenburg 2020; Schuilenburg 2024).

In de Nederlandse context kan hierbij worden gedacht aan het voorspellen van criminaliteit door de Nationale Politie via het Criminaliteits Anticipatie Systeem (CAS). Dataprofessionals die bij dit systeem betrokken waren, gaven aan dat de gebruikte patroonherkenningstechnieken te complex waren en het gebruik van zelflerende algoritmen hierdoor ondoorzichtig was (Waardenburg 2021). Een ander voorbeeld is de toeslagenaffaire, waarin het besluitvormingssysteem een (zelflerend) algoritme bevatte dat risicoprofielen van aanvragers van kinderopvangtoeslagen opstelde (Galič e.a. 2023). Mede op basis van AI werden diverse ouders door de Belastingdienst onterecht voor fraudeur aangezien. Zij moesten daarom ontvangen subsidies terugbetalen (met heffing) en kwamen niet in aanmerking voor nieuwe toeslagen. Nadat een gang naar de rechter uiteindelijk de onrechtmatigheid hiervan boven water kreeg, koos de regering ervoor om de stekker uit het systeem te trekken en slachtoffers financieel te compenseren.

In de huidige platformeconomie bestaat ook het risico dat AI leidt tot vormen van uitbuiting (Hiah 2023). Dat komt omdat het algoritme

16 Dit is onder meer aan de orde geweest in de discussie over het risicotaxatie-instrument OxRec, een risicotaxatie- en adviesinstrument dat de kans op recidive bij delinquenten in kaart brengt (Van Dijck 2020; Maas e.a. 2020). In tegenstelling tot zelflerende algoritmen gaat het bij de toepassing van dergelijke instrumenten om 'rule-based' algoritmen, dat wil zeggen: een reeks 'als-dan-algoritmen'. Om die reden laten we deze discussie hier verder buiten beschouwing.

17 Zie https://eucpn.org/sites/default/files/document/files/PP%20%282%29.pdf, bezocht op 19 januari 2024.

doi: 10.5553/JV/016758502024050001002

bepaalt welke diensten werknemers tegen welke vergoeding mogen draaien aan de hand van onder andere piek- en daluren, afstanden en het weer. Wie te vaak bezwaar maakt tegen het steeds veranderende schema krijgt door het systeem minpunten toegekend. Een negatieve score heeft als gevolg dat de werknemer ('zelfstandig ondernemer' in het jargon van de platformeconomie) in de nabije toekomst minder of geen werk toebedeeld krijgt en er daardoor dus financieel op achteruit zal gaan. Dit brengt deze werknemers in een situatie waarbij er sprake is van 'uit feitelijke verhoudingen voortvloeiend overwicht', een bestanddeel uit het wetsartikel mensenhandel (art. 273f van het Wetboek van Strafrecht (Sr)).

Omdat het trainen van AI-modellen gebaat is bij grote hoeveelheden data, bestaat ook het risico dat *scraping* op het internet leidt tot het overtreden van copyrightregels. Zo klaagde *The New York Times* OpenAI en Microsoft aan omdat hun chatbots zijn getraind met miljoenen geschreven artikelen van dit dagblad.[18] Ook bekende schrijvers en kunstenaars zijn op zijn zachtst gezegd niet gecharmeerd van het feit dat hun stijl door bepaalde AI-aanbieders wordt nagebootst.[19] De klagers geven aan dat zij geen toestemming hebben gegeven om hun werk als trainingstool voor chatbots of *image generators* te laten gebruiken.

Op kleinschaliger niveau kan AI zorgen voor onvoorziene delicten door het creëren van nepfeiten. Zo bleek dat Michael Cohen, de voormalige advocaat van Donald Trump, door AI gegenereerde jurisprudentie in zijn eigen rechtszaak inbracht.[20] Toen de rechter de verwijzingen niet kon terugvinden, bleek dat de gegevens door Googles chatbot Bard Gemini verzonnen waren. Door Cohen en zijn advocaat werd het afgedaan als een slordigheid en een betreurenswaardig misverstand. In een andere Amerikaanse strafzaak zijn twee advocaten veroordeeld tot het betalen van een boete van $ 5.000 omdat zij de rechtbank naar nepzaken hadden verwezen. Een van de advocaten schreef: 'I simply had no idea that ChatGPT was capable of fabricating entire case citations or judicial opinions, especially in a manner that appear-

18 Zie www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html, bezocht op 19 februari 2024.

19 Zie https://abcnews.go.com/US/famous-authors-lawsuit-chatgpt-maker-openai-begins-initial/story?id=105239215, bezocht op 19 februari 2024.

20 Zie www.theverge.com/2023/12/29/24019067/michael-cohen-former-trump-lawyer-google-bard-ai, bezocht op 19 februari 2024.

doi: 10.5553/JV/016758502024050001002

ed authentic (…). I deeply regret my decision to use ChatGPT for legal research, and it is certainly not something I will ever do again.'[21]

Conclusie en discussie

Dat AI voor het plegen van criminaliteit wordt ingezet, krijgt in het beleid en de wetenschap nog weinig aandacht. De focus ligt vooral op 'klassieke' vormen van cyber- en gedigitaliseerde criminaliteit en de achtergronden van de daders en slachtoffers hiervan. Zo zijn in het campagneprogramma van de Nederlandse overheid voor cybercrime geen thema's opgenomen die specifiek betrekking hebben op AI.[22] Om hierin verandering te brengen hebben wij een breed pallet aan actuele en relevante verschijningsvormen van AI-criminaliteit gegeven. Deze variëren van deepfakes en de automatische productie van malware tot situaties die leiden tot discriminatie, uitbuiting en algoritmische (markt)manipulatie. Het gaat ons hierbij niet om een uitputtende opsomming, maar om vormen van AI-criminaliteit die stof tot denken geven. Bij dit alles hebben wij AI-criminaliteit onderverdeeld in drie verschijningsvormen. Ten eerste gaat het om traditionele vormen van criminaliteit die met behulp van AI wordt gepleegd. Ten tweede wordt criminaliteit tegen AI gepleegd; de AI is hierbij doelwit geworden. Tot slot kan AI ook zonder menselijk handelen vormen van criminaliteit plegen. De AI wordt dan een 'dader'.

In contrast met bekende cyberdelicten als hacken kunnen bepaalde vormen van AI-criminaliteit als zeer laagdrempelig worden gezien. AI-instrumenten zoals gratis beschikbare programma's als ChatGPT bieden de mogelijkheid om het plegen van de strafbare delicten te automatiseren. Technische kennis is hiervoor in veel gevallen amper nog nodig, bijvoorbeeld voor het maken van kwaadwillende deepfakes. Je zou ook kunnen zeggen dat de vaardigheid zich hierdoor verplaatst van de menselijke dader naar AI of het LLM. Wat betekent dit voor de aard en omvang van criminaliteit? De populariteit van chatbots blijft immers toenemen. Google is uitgegroeid tot de belangrijkste poort tot het internet, maar wat als ChatGPT de nieuwe Google wordt? De Uni-

21 Zie www.businessinsider.com/chatgpt-generative-ai-law-firm-fined-fake-cases-citations-legal-2023-6?international=true&r=US&IR=T, bezocht op 19 februari 2024.

22 Zie https://open.overheid.nl/documenten/dpc-c58d4ae18985653cda270c3efac7b8099196ffb0/pdf, bezocht op 19 februari 2024.

doi: 10.5553/JV/016758502024050001002

ted Nations Office of Drugs and Crime (2024) schrijft dat er aanwijzingen zijn van een groeiend gebruik van deepfakes en andere 'malicious AI' om fraude te plegen, personen te chanteren met seksbeelden, of om zich online voor te doen als een politieagent. Daarentegen stellen Microsoft en OpenAI dat zij vooralsnog niet hebben gezien dat er sprake is van 'particularly novel or unique AI-enabled attack or abuse techniques resulting from threat actors' usage of AI'.[23]
Naast de laagdrempeligheid van AI-criminaliteit en de technisch hoge kwaliteit ervan baart de impact ervan zorgen (Küsters 2022; Daníelsson & Uthemann 2023). Zijn bepaalde AI-modellen binnen een sector leidend ('systeemtechnologie'), dan ontstaat ook het gevaar van groepsdenken en systeemrisico's. Zeker in de financiële sector kunnen AI-modellen binnen bepaalde bandbreedtes helpen bij het verantwoord nemen van beslissingen. Maar als het model gevoed wordt met vuile data, kunnen deze beslissingen verkeerd uitpakken. Bovendien is de AI getraind op wat al bestaat en bekend is. Hoe dergelijke systemen omgaan met *black swans* (uitzonderlijke situaties) is daarom uiterst onzeker (Daníelsson e.a. 2022).

In ons artikel hebben wij ingezoomd op actuele verschijningsvormen van AI-criminaliteit. Maar niet uit het oog moet worden verloren dat er ook grijze gebieden zijn. Neem bijvoorbeeld het Amerikaanse blad *Sports Illustrated*. Zonder zijn lezers hierover te informeren publiceerde dit blad verhalen die louter en alleen door AI waren geschreven. Bij de artikelen waren ook verzonnen biografieën en door AI gegenereerde foto's van zogenaamde auteurs geplaatst.[24] Toen dit naar buiten kwam, voelden de lezers zich bedrogen. Alle problemen hadden waarschijnlijk kunnen worden vermeden als de lezers van tevoren waren geïnformeerd dat de desbetreffende stukken niet door een persoon waren geschreven. Aan de andere kant, hoe kan een niet-bestaande auteur aansprakelijk worden gehouden voor de gepubliceerde content als deze schadelijke gevolgen (denk aan de aantasting van de goede naam van een persoon of een bedrijf) blijkt te hebben? Daarmee komen we aan bij juridische discussies over aansprakelijkheid, verantwoordelijkheid en strafbaarheidstelling van AI-criminaliteit. Verschillende kwalijke vormen van AI-criminaliteit kunnen met

23 Zie www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/, bezocht op 19 februari 2024.
24 Zie https://futurism.com/sports-illustrated-ai-generated-writers, bezocht op 19 februari 2024.

doi: 10.5553/JV/016758502024050001002

het huidige strafrecht worden aangepakt, door gebruik te maken van bestaande delictsomschrijvingen, waaronder afpersing, discriminatie en oplichting (Van der Sloot e.a. 2021). Wanneer een crimineel AI gebruikt om een *phishing attack* uit te voeren om inloggegevens te achterhalen, is deze persoon strafbaar omdat er sprake is van computervredebreuk en oplichting bijvoorbeeld. De discussie wordt echter ingewikkelder wanneer AI zelf criminaliteit pleegt doordat via machine learning eigenstandig stappen zijn ondernomen ('door AI'). Zo is voor aansprakelijkheid in de zin van artikel 6:162 van het Burgerlijk Wetboek (BW) de toerekenbaarheid van de daad aan de dader vereist. Maar wie is hier de dader? De persoon die het zelflerende algoritme heeft ontworpen, het bedrijf achter het AI-systeem of het algoritme zelf? Of wat als blijkt dat bepaalde data waarmee het AI-systeem is getraind niet gebruikt hadden mogen worden, bijvoorbeeld vanwege copyrightrechten of op discriminatoire gronden?
Tot slot moeten de grote financieel-economische belangen die met de ontwikkeling van AI gemoeid zijn niet uit het oog worden verloren. Zakentijdschrift *Forbes* gaf aan dat de AI-markt in 2022 $ 137 miljard waard was, met naar verwachting een exponentiële groei in de komende jaren. De ontwikkelaars die de markt weten te domineren, kunnen rekenen op een enorme omzet. Tegelijk gaan de ontwikkelingen op het gebied van AI razendsnel. Om deze reden gaven diverse grote spelers onlangs aan dat er pas op de plaats gemaakt moet worden met de ontwikkeling van AI.[25] Alleen als de risico's en effecten beter konden worden ingeschat, zouden er verdere stappen mogen worden gezet. Verschillende ondertekenaars die in de AI-industrie werkten, gingen op de achtergrond echter door met het ontwikkelen van AI-systemen, onder wie Tesla-baas Elon Musk en zijn chatbot Grok. Een cynische toeschouwer zou daarom kunnen denken dat de oproep een poging was om de competitie een hak te zetten.[26] Je zou ook kunnen zeggen dat het pleidooi voor een tijdelijke stop op de ontwikkeling van nieuwe AI-systemen de aandacht afleidt van de concrete problemen in het heden, waaronder manipulatie en discriminatie door AI-criminaliteit. AI-criminaliteit is immers 'here to stay' – en gaat alleen maar toenemen in omvang en complexiteit.

25 Zie https://futureoflife.org/open-letter/pause-giant-ai-experiments/, bezocht op 12 januari 2024.
26 Zie www.businessinsider.com/openai-elon-musk-pause-development-letter-never-happen-2023-4?international=true&r=US&IR=T, bezocht op 12 januari 2024.

doi: 10.5553/JV/016758502024050001002

Literatuur

Blauth e.a. 2022
T.F. Blauth, O.J. Gstrein & A. Zwitter, 'Artificial Intelligence Crime: an overview of malicious use and abuse of AI', *IEEE Access* (10) 2022, p. 77110-77122.

Boshmaf e.a. 2013
Y. Boshmaf, I. Muslukhov, K. Beznosov & M. Ripeanu, 'Design and analysis of a social botnet', *Computer Networks* (57) 2013, afl. 2, p. 556-578.

Brundage e.a. 2018
M. Brundage e.a., *The malicious use of artificial intelligence: forecasting, prevention, and mitigation*, Apollo University of Cambridge Repository 2018, https://doi.org/10.17863/CAM.22520.

Caldwell e.a. 2020
M.L. Caldwell, J.T.A. Andrews, T. Tanay & L.D. Griffin, 'AI-enabled future crime', *Crime Science* (9) 2020, afl. 1, p. 1-13.

CBS 2023
CBS (Centraal Bureau voor de Statistiek), *Online veiligheid en criminaliteit 2022*, Den Haag 2023.

Custers 2021
B.H.M. Custers, 'Artificiële intelligentie in het strafrecht: een overzicht van actuele ontwikkelingen', *Computerrecht* 2021, afl. 4, p. 330-339.

Daníelsson & Uthemann 2023
J. Daníelsson & A. Uthemann, 'On the use of artificial intelligence in financial regulations and the impact on financial stability', 2023, https://ssrn.com/abstract=4604628.

Daníelsson e.a. 2022
J. Daníelsson, R. Macrae & A. Uthemann, 'Artificial intelligence and systemic risk', *Journal of Banking & Finance* (140) 2022, https://doi.org/10.1016/j.jbankfin.2021.106290.

Das & Schuilenburg 2020
A. Das & M. Schuilenburg, 'Garbage in, garbage out. Over predictive policing en vuile data', *Beleid en Maatschappij* (47) 2020, afl. 3, p. 254-268.

Van Dijck 2020
G. van Dijck, 'Algoritmische risicotaxatie van recidive. Over de Oxford Risk of Recidivism tool (OXREC), ongelijke behandeling en discriminatie in strafzaken', *Nederlands Juristenblad* (95) 2020, afl. 25, p. 1784-1790.

Europol 2023
Europol, *ChatGPT. The impact of Large Language Models on law enforcement* (Tech Watch Flash), Luxemburg: Publications Office of the European Union 2023.

doi: 10.5553/JV/016758502024050001002

Faassen e.a. 2021
J.N. Faassen, J. Reef & M.J.F. van der Wolf, 'Virtuele kinderpornografie als behandelinstrument in de forensische psychiatrie: een Catch-22. Verkenning van de gedragskundige en juridische mogelijkheden', in: J. Altena, J. Cnossen, J. Crijns, P. Schuyt & J. ten Voorde (red.), *In onderlinge samenhang. Liber amicorum Tineke Cleiren*, Den Haag: Boom juridisch 2021, p. 319-335.

Galič e.a. 2023
M. Galič, A. Das & M. Schuilenburg, 'AI and administration of criminal justice. Report on the Netherlands', *e-Revue Internationale de Droit Pénal* 2023, p. 5-61.

Hannigan e.a. 2023
T. Hannigan, I.P. McCarthy & A. Spicer, 'Beware of botshit: how to manage the epistemic risks of generative chatbots', *Business Horizons* 2023, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4678265.

Hayward & Maas 2020
K. Hayward & M.M. Maas, 'Artificial intelligence and crime: a primer for criminologists', *Crime, Media, Culture* (17) 2020, afl. 2, p. 209-233.

Hiah 2023
J. Hiah, 'Surveillance en controle als work based harms', *Tijdschrift over Cultuur & Criminaliteit* (13) 2023, afl. 2, p. 13-32.

Horowitz 2019
M.C. Horowitz, 'When speed kills: lethal autonomous weapon systems, deterrence and stability', *Journal of Strategic Studies* (42) 2019, afl. 6, p. 764-788.

Jeong 2020
D. Jeong, 'Artificial intelligence security threat, crime, and forensics: taxonomy and open issues', *IEEE Access* (8) 2020, p. 184560-184574.

King e.a. 2020
T.C. King, N. Aggarwal, M. Taddeo & L. Floridi, 'Artificial Intelligence Crime: an interdisciplinary analysis of foreseeable threats and solutions', *Science and Engineering Ethics* (26) 2020, afl. 1, p. 89-120.

Küsters 2022
A. Küsters, *AI as systemic risk in a polycrisis. The danger of algorithmic prediction in unknown environments* (cepAdhoc No 15), Centrum für Europäische Politik 2022.

Maas e.a. 2020
M. Maas, E. Legters & S. Fazel, 'Professional en risicotaxatie-instrument hand in hand', *Nederlands Juristenblad* (95) 2020, afl. 28, p. 2055-2060.

doi: 10.5553/JV/016758502024050001002

Mai e.a. 2023
K.T. Mai, S. Bray, T. Davies & L.D. Griffin, 'Warning: humans cannot reliably detect speech deepfakes', *PLoS ONE* (18) 2023, afl. 8, https://doi.org/10.1371/journal.pone.0285333.

Maras & Alexandrou 2018
M. Maras & A. Alexandrou, 'Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos', *International Journal of Evidence and Proof* (23) 2018, afl. 3, p. 255-262.

National Cyber Security Centre 2024
National Cyber Security Centre, *The near-term impact of AI on the cyber threat*, 2024, www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat.

Oerlemans & Van der Wagen 2021
J-J. Oerlemans & W. van der Wagen, 'Types of cybercrime and their criminalisation', in: W. van der Wagen, J-J. Oerlemans & M. Weulen Kranenbarg (red.), *Essentials in cybercrime. A criminological overview for education and practice*, Den Haag: Eleven International Publishing 2021, p. 53-97.

Peeters & Schuilenburg 2018
R. Peeters & M. Schuilenburg, 'Machine justice: governing security through the bureaucracy of algorithms', *Information Polity* (23) 2018, afl. 3, p. 267-280.

Ratner 2021
C. Ratner, 'When "Sweetie" is not so sweet: artificial intelligence and its implications for child pornography', *Family Court Review* (59) 2021, afl. 2, p. 386-401.

Richardson e.a. 2019
R. Richardson, J. Schultz & K. Crawford, 'Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice', *New York University Law Review Online* (94) 2019, afl. 192, p. 192-233.

Scharre 2015
P. Scharre, 'Counter-swarm: a guide to defeating robotic swarms', *War on the Rocks* 31 maart 2015, https://warontherocks.com/2015/03/counter-swarm-a-guide-to-defeating-robotic-swarms/.

Scharre 2016
P. Scharre, 'Autonomous weapons and operational risk', *Center for a New American Security* 29 februari 2016, www.cnas.org/publications/reports/autonomous-weapons-and-operational-risk.

doi: 10.5553/JV/016758502024050001002

Schermer 2022
B.W. Schermer, *De gespannen relatie tussen privacy en cybercrime* (oratie Leiden), 2022.
Schuilenburg 2024
M. Schuilenburg, *Making surveillance public. Why you should be more woke about AI and algorithms*, Den Haag: Boom 2024.
Schuilenburg & Soudijn 2023
M. Schuilenburg & M. Soudijn, 'Big data policing: the use of big data and algorithms by the Netherlands police', *Policing: A Journal of Policy and Practice* (17) 2023, https://doi.org/10.1093/police/paad061.
Seymour & Tully 2016
J. Seymour & P. Tully, *Weaponizing data science for social engineering: automated E2E spear phishing on Twitter*, 2016, www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf.
Slobogin & Brayne 2022
C. Slobogin & S. Brayne, 'Surveillance technologies and constitutional law', *Annual Review of Criminology* (6) 2022, p. 1-22.
Van der Sloot e.a. 2021
B. van der Sloot, Y. Wagensveld & B-J. Koops, Deepfakes. *De juridische uitdagingen van een synthetische samenleving*, Tilburg University 2021.
United Nations Office of Drugs and Crime 2024
United Nations Office of Drugs and Crime, *Casinos, money laundering, underground banking, and transnational organized crime in East and Southeast Asia: a hidden and accelerating threat* (Technical Policy Brief), 2024.
Waardenburg 2021
L. Waardenburg, *Behind the scenes of artificial intelligence. Studying how organizations cope with machine learning in practice*, Alblasserdam: Haveka 2021.
WRR 2023
WRR (Wetenschappelijke Raad voor het Regeringsbeleid), *Opgave AI. De nieuwe systeemtechnologie*, Den Haag 2023.
Zuboff 2019
S. Zuboff, *The age of surveillance capitalism. The fight for a human future at the new frontier of power*, New York, NY: Public Affairs 2019.

doi: 10.5553/JV/016758502024050001002