

*Pieter Tops, Jonas Stuurman, Jimmy Maan, Sven Janssen, Derkjan Elzinga
& Willem-Jan van den Heuvel*

Dit boek geeft een caleidoscopisch overzicht van de mogelijkheden die de ontwikkelingen op het gebied van datawetenschap en artificiële intelligentie (AI) bieden bij het analyseren en terugdringen van ondermijning en georganiseerde misdaad in Nederland.

We willen met dit boek laten zien welke diversiteit aan gebruiksmogelijkheden van datawetenschap en AI er op dit terrein is. Die variëren van heel concreet (analyse van rechtbankuitspraken) tot heel abstract (versterking van de veiligheid van havens via een smart container) en van onmiddellijke toepasbaarheid (voorwaarden voor 'datavolwassenheid' in gemeenten) tot verder weg liggende potentie (gebruik van datawetenschap om de effecten van drugslegalisering in kaart te brengen).

Het boek biedt geen systematisch overzicht van alle ontwikkelingen, maar laat zien wat er al gebeurt en welke mogelijkheden er zijn; in die zin wil het eerder een inspiratiebron zijn dan een handboek. De hoofdstukken wijzen waar nodig op de problemen en gevaren die met de ontwikkeling van datawetenschap en AI kunnen samenhangen, maar dit is geen boek dat de wereld daarvoor nog een keer wil waarschuwen. Het wil eerder mogelijkheden en potentie beschrijven, en de voor- en nadelen daarvan laten zien.

Uitgangspunt van de analyses is geweest om zo veel mogelijk aan te sluiten bij praktische vragen en problemen in plaats van de beschikbaarheid van data als uitgangspunt te nemen. Er zijn overigens wel steeds meer data beschikbaar, zoals in verschillende hoofdstukken in dit boek zal blijken, en dat is weer een stimulans voor de verdere ontwikkeling en toepassing van datawetenschap op het genoemde onderwerp.

1 Dit hoofdstuk is deels gebaseerd op de oratie van Tops (2022).

Wat is datawetenschap?

Datawetenschap is gericht op het ontwikkelen van betekenisvolle en actiegerichte inzichten uit grote hoeveelheden data. Daartoe worden wiskunde en statistiek samengebracht met artificiële intelligentie, computer- en informatiewetenschap en met specifieke materie- of domeinkennis. Bedoeling is om uit die omvangrijke hoeveelheden data nieuwe kennis – of nieuwe waarde – te ontwikkelen en daarmee besluitvorming te verbeteren (Provost & Fawcett, 2013, p. 5). Men kan hierbij drie essentiële deelgebieden duiden: data-engineering (het identificeren, cureren, integreren en ordenen van data), data-analyse (het ontleden van data ten behoeve van beschrijvende, voorspellende en voorschrijvende analyses) en databesluitvorming en innovatie (het omzetten van data-analyses in concrete inzichten, innovaties en waardevolle toepassingen) (De Maat, 2021, p. 3).

Datawetenschap is mogelijk gemaakt door drie min of meer gelijktijdige ontwikkelingen, die met name sinds het begin van deze eeuw een grote vlucht genomen hebben (De Maat, 2021, WRR, 2021). In de eerste plaats het beschikbaar komen van enorm hoeveelheden data, ook wel big data genoemd, met name als gevolg van de stormachtige ontwikkeling van internet en sociale media. Het is niet overdreven om te spreken van een 'dataficering' en in tandem een 'digitalisering' van onze samenleving, waarin data beschikbaar zijn over zo'n beetje alle activiteiten en handelingen die door mensen verricht worden. Ook de aanwezigheid en plaatsing van sensoren, zoals camera's, maar ook geluids- of reukdetectoren, spelen daar een cruciale rol in. Hierbij lopen de fysieke en digitale wereld in toenemende mate in elkaar over.

In de tweede plaats groeit niet alleen de hoeveelheid data, maar ook de verscheidenheid ervan. Alles is data geworden. Niet alleen beelden zoals die via foto's, filmpjes en camera's beschikbaar komen, maar ook tekstberichten die we elkaar via mobiele telefoons en computers toezenden. Er is sprake van gestructureerde, ongestructureerde en semigestructureerde data. Voor de mate waarin ze bewerkt en gebruikt kunnen worden als bronnen van kennis en waarde maakt de mate van gestructureerdheid eigenlijk steeds minder uit. Er zijn schattingen dat ongeveer 80% van de beschikbare data ongestructureerd zijn (WRR, 2021).

In de derde plaats is de snelheid waarmee data beschikbaar komen en bewerkt kunnen worden enorm gegroeid. Steeds meer data zijn in real-time beschikbaar om te bewerken en te interpreteren.

Snelheid, hoeveelheid en diversiteit (in het Engels *velocity*, *volume* en *variety*) staan wel bekend als drie oorspronkelijke V's van big data. Ze zijn later aangevuld met allerlei andere v's waarvan *veracity* en *value* de belangrijkste zijn. Veracity gaat over de kwaliteit van de beschikbare data, dat wil zeggen over de mate waarin de gegevens juist en actueel zijn; geen onbelangrijk punt als er in toenemende mate

beslissingen op gebaseerd worden. Value heeft betrekking op de waarde die een organisatie uit de beschikbare data weet te halen. Zijn het data waar je wat aan hebt voor het functioneren of de doorontwikkeling van je organisatie? Daarbij gaat het niet alleen om het efficiënter maken van de bestaande organisatie, maar ook om het vernieuwen en transformeren daarvan. In dit laatste geval wordt wel gesproken van datagedreven organisatiemodellen, die een disruptieve werking zouden hebben. We zullen daar straks een paar voorbeelden van geven.

Nauw verbonden met de ontwikkeling van datawetenschap, en in zekere zin een onmisbare voorwaarde daarvoor, is de ontwikkeling van artificiële intelligentie geweest. Volgens de AI High Level Expert Group van de Europese Unie (EU) heeft AI betrekking op systemen die intelligent gedrag vertonen door zelf hun omgeving te analyseren en die met enige graad van autonomie actie kunnen ondernemen om specifieke doelen te bereiken (HLEG, 2019).

Cruciaal is hier natuurlijk die ‘enige graad van autonomie’. Dit betekent dat de systemen niet alleen zelf kunnen leren, maar dat zij het geleerde ook zonder menselijke tussenkomst in de praktijk zouden kunnen brengen. Dit klinkt dapper en interessant, maar van daadwerkelijke nabootsing van menselijk intellect zijn we nog ver verwijderd. De paradox van Moravec is bepaald nog niet opgeheven: activiteiten die voor mensen heel moeilijk zijn, kunnen gemakkelijk door computers worden uitgevoerd (zoals complexe berekeningen of schaken). Maar activiteiten die voor mensen relatief makkelijk zijn, blijken voor computers vaak weer heel ingewikkeld (zoals de correcte perceptie van objecten, het ontwikkelen van creativiteit of het maken van morele afwegingen) (WRR, 2021, p. 45).

Dit neemt niet weg dat net als de datawetenschap ook de ontwikkeling van AI de laatste jaren in een stroomversnelling is gekomen. Schematisch kan die ontwikkeling worden weergegeven als die van expertsystemen naar *machine leren* en vervolgens naar *diep leren*. In werkelijkheid is de ontwikkeling niet zo schematisch en chronologisch geweest, maar heeft deze naast en door elkaar plaatsgevonden. Toch schetsen de verschillende stadia de ontwikkeling naar complexere vormen van AI.

Expertsystemen zijn een vorm van ‘regelgeleid leren’: de computer is dan een lerend object door logische regels te coderen met formules van het type ‘als x dan y’. Die regels zijn gebaseerd op de kennis en ervaring van menselijke experts. De computer voert die kennis en ervaringen in op basis van tevoren opgestelde regels. Zij kunnen leren met of zonder supervisie.

In machine leren kunnen artificiële neurale netwerken een belangrijke rol spelen, naast andere instrumenten zoals *Support Vector Machines*, *Decision Trees* en dimensionaliteitsreductietechnieken. Neurale netwerken hebben de afgelopen jaren een

enorme vlucht genomen en proberen de werking van neuronen in het menselijke brein te simuleren. Kunstmatige neuronen worden gevoed met grote hoeveelheden data om daar zelf patronen uit te zien te destilleren.

De ontwikkeling van artificiële neurale netwerken maakt een enorm snelle ontwikkeling door waarbij er steeds meer lagen van neuronen aan modellen worden toegevoegd waardoor zij steeds ‘slimmer’ worden. Inmiddels spreken we ook van diep leren (*deep learning*). Daarbij wordt gebruikgemaakt van neurale netwerken met ‘diepere’ lagen, waarbij in elke laag aanvullend wordt geleerd.

De variant van diep leren kan zelfstandig leren van eigen fouten en die fouten herstellen, het kan beter omgaan met uitzonderingen en het kan abstractere kenmerken ontdekken. Maar diep leren kan ook leiden tot diepe misstanden, bijvoorbeeld als het wordt ingezet om ‘deepfakes’ te creëren (Van der Sloot e.a., 2022).

Sinds de introductie van nieuwe generatieve AI in de herfst van 2022, lijkt de samenleving ‘betoverd’ door haar schier onmetelijke vermogen nieuwe teksten, muziek, foto’s en filmpjes te genereren op basis van vooraf voorgeschotelde (enorme aantallen) voorbeelden. Generatieve AI is een specifieke tak van AI die niet enkel in staat is om dingen te herkennen, te correleren, clusteren en classificeren, maar ook om zelf voorbeelden te creëren. Hierbij maakt zij doorgaans gebruik van een specifieke categorie dieplerentechnieken genaamd *Generatieve Pre-trained Transformers* (GPT’s); een uitvinding van het Californische bedrijf Open-AI dat achter producten zoals ChatGPT zit. Simpel gezegd zijn GPT’s gebaseerd op een *transformer architecture*, een speciaal type van dieplerennetwerk, dat zelfattentie (*self-attention*) implementeert. Zelfattentie stelt de GPT’s in staat om relaties tussen ‘tokens’ (bijv. woorden) te leggen ongeacht waar ze in een inputdataset staan, en kunnen zodoende verder kijken dan andere dieplerennetwerken die meestal enkel aanpalende tokens (woorden) meenemen in hun analyse.

Ten tijde van het schrijven van dit boek worden aan GPT’s bijna goddelijke kwaliteiten toegezegd. Dat gezegd hebbende, verwachten wij dat GPT’s in de korte tot middellange termijn hun toepassing zullen vinden in technologieën zoals chatbots (bijv. om schadelijke effecten van ondermijning die online gemeld worden te verzamelen), virtuele assistenten (bijv. bij het assisteren in het onderzoeken van vermeende vastgoedfraude), en naturallanguageprocessinganalyses (bijv. voor het analyseren van cryptochats).

Artificiële intelligentie wordt inmiddels overigens al op vele manieren toegepast in de ontwikkeling van voorspellingssystemen (*predictive analytics* en *predictive policing*), in beeldherkenning (*computer vision*), in taalverwerkingssystemen (*natural language processing*), in spraakherkenning (*speech recognition*) en in robotica, waarin alle

soorten van AI samenkomen en worden gecombineerd met robots, die in de fysieke wereld taken kunnen uitvoeren. Een voorbeeld hiervan is de slimme container die we in hoofdstuk 8 zullen behandelen.

Door AI gestuurde machines nemen beslissingen die serieuze consequenties kunnen hebben voor personen en samenlevingen. Bijvoorbeeld of iemand wel of niet in aanmerking komt voor een lening of hypotheek of wel of niet extra wordt gecontroleerd door de belastingdienst of de bank. De werking van die autonome algoritmes is vaak zo ondoorzichtig dat het vrijwel onmogelijk is om te achterhalen hoe het systeem tot een beslissing is gekomen. AI is dan een soort van *black box* worden; er zijn wel resultaten, maar waar ze vandaan komen en waar ze op gebaseerd zijn, is betrekkelijk onduidelijk. Bovendien kunnen al dan niet subtiele vooroordelen in het systeem geslopen zijn. Is die artificiële intelligentie eigenlijk wel te vertrouwen, is dan de prangende vraag. Vandaar dat hun toepassing met serieuze waarborgen dient te worden omgeven. We komen daar in verschillende hoofdstukken op terug.

Wat is ondermijning?

Ondermijning is te omschrijven als de schadelijke maatschappelijke effecten van georganiseerde misdaad. Zij tast daarmee de beoogde en legale werking van de maatschappij aan. Dat is niet altijd de expliciete bedoeling van criminelen, maar is doorgaans wel een effect van hun handelen (Van der Steen e.a., 2016). Georganiseerde misdaad onttrekt zich aan normen die in de reguliere maatschappij gelden, oefent aantrekkingskracht uit via de – vaker valse dan realistische – belofte van geld en status, en gebruikt de reguliere maatschappij voor haar activiteiten.

Wij maken een analytisch onderscheid tussen georganiseerde misdaad en ondermijning; we trekken de begrippen uit elkaar (zie ook Tops en Van der Torre, 2024). Georganiseerde misdaad heeft betrekking op criminele samenwerkingsverbanden: hoe zij ontstaan, functioneren en bestreden kunnen worden. Ondermijning heeft betrekking op de schadelijke effecten van georganiseerde misdaad op de samenleving. Georganiseerde misdaad is een criminologisch begrip, ondermijning een bestuurskundig. We gebruiken het om die schadelijke effecten te ordenen en nader te analyseren.

Het onderscheid tussen georganiseerde misdaad en ondermijning heeft praktische betekenis. Het leidt tot rolverduidelijking. De aanpak van georganiseerde misdaad als zodanig is primair een taak van de opsporingsinstanties: politie, Openbaar Ministerie (OM) en bijzondere opsporingsdiensten, zoals de Fiscale Inlichtingen- en Opsporingsdienst (FIOD), de Nederlandse Arbeidsinspectie en de Nederlandse Voedsel- en Warenautoriteit (NVWA). De aanpak van ondermijning daarentegen is primair een verantwoordelijkheid voor het openbaar bestuur, waarin gemeenten een cruciale taak hebben, met een hoofdrol voor de burgemeester. Tegelijkertijd is er