HANDBOEK · EDITIE 2026

AI-Ethiek

Principes, praktijk en de Europese context

EEN SYSTEMATISCH ONDERZOEK NAAR DE ETHISCHE EN MAATSCHAPPELIJKE VRAAGSTUKKEN ROND KUNSTMATIGE INTELLIGENTIE.

AUTEUR

Elise Konings

EDITIE

MMXXVI

HANDBOEK

AI-Ethiek

Principes, praktijk en de Europese context

C O L O F O N

AI-Ethiek · Handboek

Principes, praktijk en de Europese context

AUTEUR Elise Konings

EERSTE EDITIE 2026

Dit handboek biedt een systematisch overzicht van de ethische en maatschappelijke vraagstukken rond kunstmatige intelligentie. Het werk is opgebouwd rond tien hoofdstukken, verdeeld over twee delen, en sluit af met een uitgebreid bronnenapparaat.

De tekst richt zich op studenten, professionals en beleidsmakers die de verhouding tussen kunstmatige intelligentie en samenleving grondig willen begrijpen. Casussen en kerninzichten verbinden abstracte principes met de dagelijkse praktijk.

[Een aantal Engelse vaktermen, zoals *bias*, *fairness*, *accountability*, *explainability*, *surveillance*, *governance*, *deployment* en *human in the loop*, zijn bewust niet vertaald. Ze functioneren als gevestigde begrippen in het internationale onderzoeks- en beleidsdebat, en een Nederlandse omschrijving zou hun precieze betekenis en draagwijdte verzwakken.]

Inhoud

DEEL I

Ethische principes

HOOFDSTUK 1

Introductie

Een neutraal systeem bestaat niet

Een neutraal systeem bestaat niet

Eeuwenlang gold menselijk oordeel als onvermijdelijk bevooroordeeld, gebonden aan stemming, status en eigenbelang. De komst van geautomatiseerde besluitvorming leek daar een uitweg uit te bieden. Data zouden doen wat mensen niet konden: consequent, onpartijdig, schaalbaar. Wat die systemen in de praktijk deden, vertelt een ander verhaal.

AI-systemen nemen steeds vaker beslissingen die het leven van mensen concreet bepalen: welke sollicitant wordt uitgenodigd, welk krediet wordt toegekend, welke diagnose wordt voorgesteld, welk nieuwsbericht als eerste verschijnt. Die beslissingen worden doorgaans niet gemotiveerd, nauwelijks betwist en zelden begrepen door de mensen die eraan onderworpen zijn.

De centrale vraag van AI-ethiek is niet of AI slim genoeg is. Zij is wat AI doet met macht, kansen en informatie, en wie daarvoor verantwoordelijk kan worden gesteld.

Wat AI is, en wat niet

AI is geen scherp afgelijnd begrip. De meeste systemen die vandaag onder die noemer circuleren zijn *narrow AI*: systemen die ontworpen zijn voor een beperkte, afgebakende taak. Ze kunnen daarbinnen indrukwekkend presteren, maar begrijpen niet wat ze doen. Buiten hun specifieke domein kunnen ze niets.

General AI, systemen die menselijk denken in brede zin benaderen, flexibel redeneren en zelfstandig nieuwe problemen aanpakken, bestaat vandaag niet. Ze blijven een onderzoeksdoel en een bron van debat, terwijl publieke discussies soms anders doen vermoeden.

Er is ook een paradox. Zodra een AI-systeem ingeburgerd raakt, noemen mensen het geen AI meer. Het wordt gewoon technologie. Dat gold voor spamfilters, voor aanbevelingssystemen, voor automatische vertaling. Nu verschuift het beeld alweer naar wat nieuwer of onbegrijpelijker lijkt. Die voortdurende horizon maakt het moeilijk om een stabiel oordeel te vormen over wat AI werkelijk kan.

Patroonherkenning, geen begrip

Wat AI-systemen doen, is patronen herkennen in grote hoeveelheden data en op basis daarvan voorspellingen doen. Een taalmodel genereert tekst door te voorspellen welk woord waarschijnlijk volgt; het begrijpt niet wat de woorden betekenen. Een beeldherkenningssysteem classificeert een foto op basis van statistische gelijkenissen met eerder gelabelde beelden; het ziet niet wat er werkelijk op staat.

Systemen die patronen herkennen in historische data lopen het risico historische ongelijkheden te reproduceren en te versterken, niet door kwade bedoelingen, maar structureel, als gevolg van de data waarop ze zijn getraind. Een systeem dat aangeleerd wordt op aanbevelingen van recruiters uit het verleden, leert ook de vooroordelen van die recruiters. Systematisch en op grote schaal.

Dat die beperkingen op het hoogste wetenschappelijke niveau ernstig worden genomen, blijkt uit een artikel in *Science* (2024) van een groep toonaangevende AI-onderzoekers, waaronder Yoshua Bengio en Geoffrey Hinton. Zij stellen dat de capaciteiten van AI-systemen snel groeien, maar dat de bijbehorende

veiligheidsgaranties en governancestructuren achterblijven. De combinatie van toenemende autonomie en beperkte controlemechanismen maakt ethische reflectie urgent.

Een korte geschiedenis van grote verwachtingen

AI is geen recent fenomeen. Alan Turing formuleerde in 1950 al de vraag of machines kunnen denken. John McCarthy gaf het onderzoeksveld zijn naam en structuur in de jaren vijftig. Die vroege geschiedenis toont iets duurzaams: AI gaat van oudsher gepaard met grote verwachtingen.

In de jaren zeventig en opnieuw in de jaren negentig zakte de interesse weg omdat systemen de beloften niet konden inlossen, de zogenaamde AI-winters. Wat de huidige golf onderscheidt, is niet zozeer de intelligentie van de systemen, maar de rekenkracht, de hoeveelheid beschikbare data en de commerciële schaal waarop ze worden ingezet. Hypegolven zijn van alle tijden; de kritische blik blijft even noodzakelijk.

Ethiek als noodzakelijke laag

Zodra AI-systemen beslissingen ondersteunen of automatiseren, raken ze aan waarden, rechten en maatschappelijke verwachtingen. Dan volstaat de vraag of een systeem snel of accuraat werkt niet meer. De relevante vragen zijn: welke waarden zitten erin vervat, wie draagt de gevolgen, en wie kan verantwoordelijk worden gesteld?

Die vraag stelt zich voor elk domein waar AI-systemen in gebruik zijn. Wanneer een AI-assistent medische symptomen beoordeelt en een onjuist advies geeft, is de verantwoordelijkheidsvraag even pertinent als in discriminatiecasussen. Het systeem genereert het advies in goed vertrouwen, op basis van patronen in zijn trainingsdata, terwijl die data overwegend bestaan uit Engelstalige medische literatuur die lokale behandelprotocollen, medicijnnamen of diagnostische conventies niet weerspiegelt. De schade is niet minder reëel omdat ze niemand persoonlijk viseerde. En wanneer de verantwoordelijkheid verdeeld is over de ontwikkelaar, de deployer en de gebruiker, dreigt ze bij niemand effectief te berusten.

Drie klassieke filosofische tradities helpen om die vragen systematisch te stellen. *Utilitarisme* focust op gevolgen: welke keuze maximaliseert welzijn voor het grootste aantal mensen? Dat is aantrekkelijk, maar complexer zodra belangen botsen of gevolgen moeilijk te meten zijn. *Deontologische ethiek* legt de nadruk op plichten en rechten: niet alles is geoorloofd, zelfs niet wanneer de uitkomst gunstig lijkt. *Virtue ethics* kijkt naar karakter en intentie: kan een systeem ontworpen worden dat eerlijk, voorzichtig of zorgzaam handelt, en wat betekent dat als een machine geen morele actor is?

Een vierde traditie, die bijzonder relevant is voor AI-systemen die over grote groepen mensen beslissen, is die van John Rawls. In *A Theory of Justice* (1971) introduceerde Rawls het gedachte-experiment van de *sluier van onwetendheid*: welke regels zouden mensen kiezen als ze niet wisten welke positie ze zelf innemen in de samenleving, arm of rijk, meerderheid of minderheid, gezond of kwetsbaar? Westerstrand (2024) werkte in *Science and Engineering Ethics* uit hoe dat kader rechtstreeks toepasbaar is op AI-ethiek: een rechtvaardig systeem is een systeem dat ook degenen die er het meest kwetsbaar voor zijn, als eerlijk zouden erkennen.

Een ander perspectief komt uit de feminist AI ethics. Virginia Eubanks toont in *Automating Inequality* (2018) hoe geautomatiseerde systemen de sociale controle over kwetsbare groepen versterken: mensen die afhankelijk zijn van publieke diensten worden aan preciezere, indringendere surveillance onderworpen dan wie het zich kan veroorloven buiten het systeem te blijven. Safiya Umoja Noble analyseert in *Algorithms of Oppression* (2018) hoe zoekmachines en aanbevelingssystemen racistische en seksistische patronen reproduceren en zo als neutraal gepresenteerde technologie ideologische effecten heeft. Beide werken tonen dat schade door AI niet toevallig is maar structureel: ze treft de mensen die al de minste macht hebben.

De capabilities approach van Martha Nussbaum (2011) voegt een derde perspectief toe. Nussbaum stelt dat rechtvaardigheid niet gaat over de maximalisatie van nut of de gelijke verdeling van middelen, maar over de vraag of mensen de fundamentele capaciteiten kunnen ontwikkelen die een waardig leven mogelijk maken. Voor AI-systemen betekent dat: een systeem is ethisch problematisch wanneer het de reële mogelijkheden van mensen inperkt, hun toegang tot werk, zorg, onderwijs of justitie, ook als het technisch correct functioneert.

Value-sensitive design, ontwikkeld door Batya Friedman en David Hendry (2019), vertrekt van een andere invalshoek: ethische waarden moeten worden ingebouwd vanaf de ontwerptafel, niet achteraf worden toegevoegd als correctie. Dat vraagt een expliciete analyse van welke waarden een systeem draagt, welke conflicten daartussen bestaan, en hoe die worden opgelost voor specifieke stakeholders.

Op Europees beleidsniveau formuleerde Luciano Floridi in opdracht van de Europese Commissie vijf kernprincipes: beneficence, non-maleficence, autonomy, justice en explicability. De UNESCO Recommendation on the Ethics of Artificial Intelligence (2021), aanvaard door alle 194 lidstaten, werkt langs vergelijkbare lijnen. Dat zulke kaders nu op internationaal niveau bestaan, toont hoe AI-ethiek geëvolueerd is van academisch debat naar juridisch en politiek beleidsinstrument.

De EU AI Act als beleidsinstrument

De Europese Artificial Intelligence Act, in werking getreden op 1 augustus 2024, is de eerste bindende regulering van AI-systemen ter wereld. Ze hanteert een risicogebaseerde aanpak: hoe hoger het risico dat een systeem stelt voor de rechten of veiligheid van mensen, hoe strenger de verplichtingen.

De implementatie verloopt gefaseerd. Vanaf 2 februari 2025 zijn de meest ingrijpende verboden van kracht: manipulatieve AI die gericht werkt op het uitbuiten van kwetsbaarheden, realtime biometrische massa-identificatie in openbare ruimten en sociale scoresystemen die burgers in hun maatschappelijke rechten beperken. Vanaf augustus 2025 gelden de regels voor general-purpose AI-modellen. Het grootste deel van de verordening, de verplichtingen voor hoog-risico toepassingen in domeinen zoals kredietverlening, aanwerving, onderwijs en strafrecht, treedt in werking in augustus 2026. De volledige toepassing voor alle systemen is voorzien voor augustus 2027.

De EU AI Act is juridisch bindend voor alle AI-systemen die op de Europese markt worden aangeboden, ook voor systemen ontwikkeld buiten Europa. Ze verplicht aanbieders van hoog-risico systemen tot technische documentatie, transparantie richting toezichthouders en, in sommige gevallen, menselijk toezicht op geautomatiseerde beslissingen.

Europese soevereiniteit en de grenzen van regulering

Regelgeving heeft echter structurele grenzen. Een van de meest tastbare is de afhankelijkheid van Europese overheden en bedrijven van niet-Europese cloudinfrastructuur. Overheden die hun bevolking beschermen via de EU AI Act of de GDPR kunnen tegelijk afhankelijk zijn van infrastructuur waarover een buitenlandse mogendheid juridisch toegang kan opeisen. Digitale soevereiniteit, controle over data, algoritmen en infrastructuur, is daarom een terugkerend thema in de Europese beleidsdiscussie dat in dit handboek op meerdere plaatsen terugkomt.

Het gevaar van ethisch theater

Toch is er reden tot scepsis over het reguleringsoptimisme. Filosoof Thomas Metzinger, voormalig lid van de High-Level Expert Group on AI van de Europese Commissie, introduceerde het begrip *ethics washing*: de praktijk waarbij bedrijven en overheden ethische discussies voeren als vervanging voor concrete regulering, of als marketinginstrument om kritiek af te wenden. Onderzoek in *AI and Ethics* (Springer, 2024) toont dat in vijf jaar tijd bijna honderd niet-bindende ethische codes werden aangenomen, terwijl de concrete effecten op de dagelijkse praktijk beperkt bleven.

Wie AI-ethiek ernstig neemt, bekijkt het ethische debat zelf met dezelfde kritische blik.

✦ KERNINZICHTEN

AI-ethiek gaat over technologie én over keuzes, waarden en gevolgen. Huidige AI-systemen zijn bijna zonder uitzondering narrow AI: sterk binnen een beperkte taak, ver verwijderd van algemene intelligentie of begrip. Omdat systemen leren van historische data, reproduceren ze ook historische ongelijkheden, systematisch en op grote schaal, zonder kwade bedoelingen. Filosofische kaders, van utilitarisme tot Rawlsiaanse rechtvaardigheid, van feminist AI ethics tot value-sensitive design, helpen om die vragen te stellen, maar zijn onvoldoende: ethische principes zonder handhaving blijven symbolisch. Regulering biedt houvast, maar de structurele afhankelijkheid van Europese instellingen van buitenlandse infrastructuur begrenst de effectiviteit van elke nationale of Europese norm. De komende hoofdstukken onderzoeken hoe die spanning zich manifesteert in concrete domeinen.

HOOFDSTUK 2

Bias

Hoe data ongelijkheid leren reproduceren

Inleiding

Data zijn geen neutrale weergave van de werkelijkheid. Ze zijn een neerslag van hoe beslissingen vroeger werden genomen, welke groepen destijds werden meegerekend, en welke niet. Een systeem dat leert van zulke data, neemt ook die vooronderstellingen over.

Dat is het vertrekpunt van algorithmische bias: niet een bewuste keuze of programmeerfouting, maar een structurele eigenschap van systemen die patronen extraheren uit een wereld die al ongelijk was.

Wat is algorithmische bias?

Algorithmische bias ontstaat wanneer een AI-systeem systematisch gunstiger of ongunstiger uitkomsten produceert voor bepaalde groepen. Het gaat dus niet om een toevallige fout, maar om een terugkerend patroon van oneerlijke behandeling. Zulke bias kan ingebouwd raken in het systeem door de trainingsdata, door keuzes in het model, door de manier waarop mensen labels toekennen, of door de context waarin een model uiteindelijk wordt ingezet.

AI-systemen spelen vandaag een rol in domeinen met grote gevolgen voor burgers. Daarom is het niet voldoende dat een systeem technisch sterk lijkt. Als een model op een systematische manier ongelijkheid versterkt, blijft het resultaat problematisch, zelfs wanneer het systeem op papier efficiënt werkt.

Waarom is bias zo belangrijk?

AI-systemen nemen vandaag mee deel aan beslissingen met directe gevolgen voor mensen: wie een lening krijgt, wie wordt uitgenodigd voor een sollicitatie, welke diagnose wordt voorgesteld, hoeveel straf een rechter oplegt. Wanneer die systemen op een systematische manier bepaalde profielen benadelen, is de schade niet abstract maar persoonlijk, en door het geautomatiseerde karakter moeilijk zichtbaar en moeilijker aan te vechten.

Bovendien versterkt bias zichzelf. Een model dat leert uit scheve historische data zal die scheeftrekking herhalen. Wanneer de output daarna opnieuw als basis dient voor nieuwe beslissingen of nieuwe data, ontstaat een feedbacklus: bias in de trainingsdata leidt tot bias in de beslissing, en die beslissing genereert nieuwe scheve data.

Waar komt bias vandaan?

Er zijn verschillende bronnen van bias. Een eerste bron is databias. Wanneer de trainingsdata een historisch scheef beeld bevatten, zal het systeem dat mee reproduceren. Een aanwervingsmodel dat leert uit cv's uit een sterk mannelijke sector kan bijvoorbeeld de impliciete voorkeur voor mannen overnemen, ook als dat nergens expliciet geprogrammeerd is.

Een tweede bron is sampling bias. Daarbij is de dataset niet representatief voor de populatie waarop het systeem later wordt toegepast. Een duidelijk voorbeeld is gezichtsherkenning die vooral getraind werd op lichte huidtypes. Wanneer zo'n systeem later gebruikt wordt bij een meer diverse populatie, stijgt de kans op fouten voor mensen die onvoldoende in de trainingsset vertegenwoordigd waren.

Hetzelfde mechanisme speelt in spraakherkenning, en raakt daarmee een bredere groep dan de termen bias en discriminatie doorgaans suggereren. Dicteersoftware en transcriptiesystemen, ook in professionele contexten zoals de gezondheidszorg of rechtbanken, zijn overwegend getraind op Standaardnederlands en Amerikaans-Engels. Regionale accenten, Vlaams, Limburgs, Zeeuws of Gronings, worden systematisch slechter herkend. In medische contexten, waar artsen dossiers inspreken, of in juridische omgevingen, waar nauwkeurigheid rechtsgevolgen heeft, kan een foutieve transcriptie directe schade veroorzaken. De bias zit niet in etnische of genderkenmerken, maar in de taalkundige samenstelling van het trainingscorpus: wie niet de standaardtaal van het dominante corpus spreekt, wordt structureel slechter bediend.

Een derde bron is model bias. Sommige modellen worden vooral geoptimaliseerd op efficiëntie of accuraatheid, zonder dat fairness even zwaar meeweegt. Daardoor kunnen modellen patronen leren die statistisch bruikbaar lijken, maar maatschappelijk onrechtvaardig zijn. Een kredietverleningsmodel kan bijvoorbeeld lage kredietwaardigheid linken aan wonen in armere buurten, en zo bestaande ongelijkheid bestendigen.

Een vierde bron is deployment bias. Zelfs wanneer een model in theorie degelijk ontworpen is, kan de concrete inzet alsnog bias introduceren. Een systeem dat getraind werd in een bepaalde context kan bij inzet in een andere context heel andere, discriminerende resultaten opleveren. De manier waarop een organisatie met een model omgaat, is dus even belangrijk als de technische bouw ervan.

Human bias in the loop en automation bias

Ook de menselijke rol in het ontstaan van bias is belangrijk. Mensen geven labels, interpreteren data en beoordelen outputs. Daardoor komt menselijke vooringenomenheid mee in het systeem terecht. Wat voor de ene beoordelaar neutraal lijkt, kan voor iemand anders positief of negatief overkomen. Dat maakt duidelijk dat bias niet alleen in code zit, maar ook in interpretatie en besluitvorming.

Er bestaat ook iets wat men automation bias noemt: de menselijke neiging om geautomatiseerde output te snel te vertrouwen. Wanneer een AI-score of voorspelling op tafel ligt, zijn mensen geneigd die als objectiever of juister te beschouwen dan een menselijke inschatting. Dat verlaagt het kritisch denkvermogen en maakt het moeilijker om foutieve of discriminerende uitkomsten nog in vraag te stellen.

Dat zie je ook in alledaagse situaties. Zodra een systeem een score, advies of rangorde geeft, ontstaat snel de indruk dat die output neutraal en betrouwbaar is. In werkelijkheid blijft ook zo'n uitkomst afhankelijk van data, ontwerpkeuzes en de context waarin het systeem gebruikt wordt.

In Groot-Brittanië werden gezondheidsnoden van vrouwen systematisch als minder zwaar beoordeeld dan die van mannen. Dat is belangrijk, omdat precies zulke evaluaties mee kunnen bepalen hoeveel zorg of ondersteuning iemand krijgt. Het voorbeeld toont hoe subtiele verschillen in AI-input kunnen uitmonden in ongelijke behandeling in de praktijk (The Guardian, 2025).

Wanneer metrics zelf bias veroorzaken

Bias ontstaat niet alleen door data of menselijke labels, maar ook door de keuze van de variabelen waarmee men iets probeert te meten. Wanneer een concept niet direct meetbaar is, grijpt men naar een *proxy*, een indirecte maatstaf die het doelconcept benadert maar er niet mee samenvalt. Proxies zijn daarbij vaak pro-

blematisch. Een bekende illustratie is het gebruik van zorgkost als proxy voor medische nood. Die twee zijn niet hetzelfde. Een groep die historisch minder toegang had tot zorg, kan lagere zorgkosten hebben, terwijl de werkelijke medische nood net hoger ligt.

Op dezelfde manier kunnen indirecte variabelen gevoelige kenmerken binnensmokkelen. Een postcode lijkt op het eerste gezicht onschuldig, maar kan in de praktijk sterk samenhangen met inkomen, etnische achtergrond, religie of toegang tot voorzieningen. Daardoor kan een model gevoelige verschillen reproduceren zonder dat die expliciet in de dataset als beschermde categorie opgenomen zijn.

Hier zit een belangrijk punt: fairness-aware variabelenselectie is noodzakelijk, maar moeilijk. In theorie klinkt het logisch dat je enkel meet wat je echt wil meten. In de praktijk vraagt dat tijd, middelen, expertise en een grondige testcultuur. Net daar wringt het vaak in organisaties.

Kredietscoremodellen voor hypotheken bieden in de Belgische en Nederlandse context een herkenbaar voorbeeld. Zulke modellen zijn doorgaans getraind op profielen van werknemers in loondienst: vaste maandelijkse stortingen, een eenduidige loonbrief, een lineaire loopbaan. Zelfstandigen en freelancers vertonen structureel andere patronen, wisselende inkomsten, meerdere rekeningen, seizoensgebonden pieken, die het model als risico-indicatoren interpreteert, ook wanneer het totale inkomen ruimschoots volstaat om de lening terug te betalen. De proxy meet iets anders dan wat bedoeld wordt: niet de werkelijke terugbetalingscapaciteit, maar de mate van gelijkenis met het loondienst-profiel waarop het model getraind is. Het gevolg is een structureel nadeel voor een demografisch breed gespreide groep, iedereen die buiten het klassieke loondienst-patroon valt.

Bias in de praktijk

Amazon stopte in 2014 een intern experiment met een AI-tool die cv's moest screenen. Het systeem had geleerd van historisch recruitmentmateriaal uit een sector die jarenlang sterk mannelijk gedomineerd was. Mannelijk gecodeerde profielen wogen daardoor zwaarder door; zelfs de vermelding van een vrouwenschaakclub kon negatief meetellen. Niemand had dat zo geprogrammeerd. Het model had het geleerd.

Loopbaanonderbrekingen kennen een vergelijkbaar mechanisme, maar treffen een veel bredere groep. Recruitment-algoritmen die filteren op loopbaancontinuïteit interpreteren periodes zonder formele tewerkstelling als risico-indicator, ongeacht of die periode te maken had met mantelzorg, ziekte, een opleiding of een ontslaggolf. Het profiel dat het systeem impliciet beloont, veronderstelt voltijdse, ononderbroken beroepsactiviteit als norm. Die norm werd nooit als expliciete ontwerpkeuze geformuleerd, maar de facto filtert ze een demografisch brede groep uit.

In de gezondheidszorg toonde een studie van Obermeyer, Powers, Vogeli en Mullainathan (2019) in Science hoe een algoritme voor de prioritering van patiëntenzorg de gezondheidsrisico's van zwarte patiënten systematisch onderschatte. De oorzaak lag in de proxy: het systeem gebruikte zorgkost als maatstaf voor medische nood. Wie historisch minder toegang had tot zorg, genereerde lagere zorgkosten, en werd daardoor als minder urgent beoordeeld. Door de proxy te vervangen door een directere maatstaf voor gezondheidsbehoeften daalde de bias met 84 procent. Bias is in veel gevallen niet onvermijdelijk, maar een gevolg van een ontwerpkeuze die ook bewust gecorrigeerd kan worden.

Gezichtsherkenning maakt de ongelijkheid in foutenmarges het meest zichtbaar. Onderzoek van Joy Buolamwini en Timnit Gebru (2018), gepresenteerd als "Gender Shades" op de FAccT-conferentie, toonde dat commerciële systemen donkergekleurde vrouwen in tot 34,7 procent van de gevallen verkeerd classificeerden, terwijl de foutmarge voor lichtgekleurde mannen slechts 0,8 procent bedroeg. De verklaring zat in de samenstelling van de trainingsdata. Wanneer zulke systemen ingezet worden in wetshandhaving, kunnen die ongelijke foutenmarges uitmonden in onterechte verdachtmakingen of arrestaties.

In mei 2025 kende een rechter voorlopige collectieve certificatie toe in de zaak Mobley v. Workday, een aanwijzing dat AI-ondersteunde sollicitatieprocessen niet alleen ethisch maar ook juridisch getoetst worden wanneer er aanwijzingen zijn van structurele benadeling (Civil Rights Litigation Clearinghouse, 2025).

De Franse gelijkheidsinstantie oordeelde in november 2025 dat het algoritme van Facebook voor jobadvertenties beroepen ongelijk distribueerde: advertenties voor mecaniciens bereikten overwegend mannen, advertenties voor kleuterleerkrachten overwegend vrouwen. Bias treft zo al wie een kans krijgt om een vacature te zien, niet alleen wie geselecteerd of afgewezen wordt (The Guardian, 2025b).

Voorbeelden uit België en Nederland

In de verzekeringssector rijzen vergelijkbare vragen. Wanneer data tonen dat alleenstaande moeders gemiddeld meer claims indienen, kan een model die groep als duurder of risicovoller behandelen, en vertaalt een kwetsbare sociale situatie zich naar een structureel financieel nadeel.

Hetzelfde geldt voor verzekeringspremies per regio. Wie in een armere wijk woont in Brussel of Rotterdam, kan te maken krijgen met hogere premies op basis van postcode, ook al zitten achter die ruwe data bijkomende factoren die het risicoprofiel vertekenen, zoals het aandeel bedrijfswagens en schadegedrag in andere inkomensgroepen. Een schijnbaar objectieve datalogica leidt zo tot oneerlijke conclusies zodra de onderliggende samenhang niet wordt geanalyseerd.

Case study: COMPAS

Een belangrijk voorbeeld is COMPAS, een systeem dat gebruikt werd om het risico op recidive in te schatten. Het werd mee ingezet in beslissingen over borgtocht, strafmaat en voorwaardelijke invrijheidstelling. Een onderzoek van ProPublica, gepubliceerd door Angwin, Larson, Mattu en Kirchner (2016) op basis van meer dan tienduizend scores uit Broward County, Florida, toonde dat zwarte beklaagden vaker als hoog risico werden gelabeld, zelfs wanneer witte beklaagden met gelijkaardige profielen en vergelijkbare hervalcijfers minder zwaar ingeschaald werden. Het valspositiefpercentage lag voor zwarte beklaagden bijna twee keer hoger dan voor witte beklaagden.

Wanneer een rechter of andere beslisser zo'n score ernstig neemt, heeft de output van het systeem directe invloed op iemands toekomst, op vrijheid of detentie, op borgstelling of weigering. Het systeem biedt de betrokkene nauwelijks houvast om die uitkomst te betwisten: de scoringslogica is ondoorzichtig en de foutmarge is per groep ongelijk.

Case study: gezichtsherkenning

In 2020 werd Robert Julian Borchak Williams onterecht gearresteerd als verdachte van een winkeldiefstal nadat een gezichtsherkenningssysteem hem foutief aan een bewakingsbeeld koppelde. Vrijheidsberoving, reputatieschade en een langdurige klachtenprocedure waren het gevolg van wat het systeem als een match had beoordeeld.

In 2026 berichtte The Guardian over Alvi Choudhury, gearresteerd voor een inbraak in een stad waar hij nooit was geweest. Dezelfde krant meldde in februari 2026 dat een klant uit een Londense Sainsbury's-vestiging werd gezet nadat personeel hem koppelde aan een Facewatch-melding. De schade door foutieve matches beperkt zich niet tot politie en justitie: ook in winkels of private beveiliging zijn de gevolgen voor de getroffene reëel, terwijl de klachtenprocedure zelden duidelijk is.

Of een systeem met sterk ongelijke foutenmarges überhaupt ingezet mag worden in wetshandhaving is geen marginale technische vraag. Fairness auditing en voorafgaande toetsing zijn essentiële voorwaarden, niet als administratieve formaliteit, maar als inhoudelijke drempel.

Feedbacklussen en versterking van bias

Wanneer een model scheve beslissingen produceert en die uitkomsten later opnieuw gebruikt worden als trainingsmateriaal of als basis voor nieuwe interventies, versterkt bias zichzelf. Bias in de data leidt tot bias in het model, vervolgens tot bias in de beslissing, en daarna opnieuw tot nieuwe scheve data, een lus zonder automatisch correctiemechanisme.

De bekendste illustratie is Microsofts chatbot Tay (2016), die leerde uit online interacties en binnen uren extreem problematisch gedrag vertoonde doordat gebruikers het systeem gerichte, racistische inputs voedden. Systemen die zonder voldoende toezicht blijven leren uit een toxische omgeving, ontsporen, en doen dat snel.

Mogelijke oplossingen

Bias auditing, het systematisch controleren van systemen op ongelijke uitkomsten, zowel voor als tijdens deployment, is een eerste technische stap. Representatievere trainingsdata verkleinen de kans dat het model een vertekend beeld van de populatie internaliseerde. Explainable AI-technieken zoals LIME en SHAP maken zichtbaar welke variabelen bijdragen aan een voorspelling, waardoor problematische patronen eerder opgespoord kunnen worden. Gespecialiseerde fairness-tools zoals Fairlearn of IBM AI Fairness 360 bieden daarvoor gestructureerde hulpmiddelen.

Technische interventies lossen het probleem echter niet volledig op. Bias is geen defect dat gerepareerd kan worden: ze is een gevolg van keuzes over data, modeldoelstellingen en inzetcontext. Governance, verantwoordelijkheidszin en een kritische testcultuur zijn minstens even onmisbaar.

Barocas, Hardt en Narayanan tonen in *Fairness and Machine Learning: Limitations and Opportunities* (MIT Press, 2023) dat fairness in machinaal leren geen wiskundig probleem is met één juiste oplossing, maar een normatieve keuze die afhangt van context, waarden en de vraag wie de gevolgen draagt. Ke-

arns en Roth betogen in *The Ethical Algorithm* (Oxford University Press, 2019) dat principes als eerlijkheid en privacy wiskundig geformaliseerd en als ontwerpkeuze in algoritmen ingebouwd kunnen worden, niet als correctie achteraf.

AI-tools kunnen ook zelf nieuwe vormen van bias introduceren. De Financial Times berichtte in december 2025 over tools die recruiters informatie laten afleiden over kandidaten, inclusief gevoelige kenmerken als zwangerschap, op basis van publiekelijk beschikbare data. Dergelijke inferenties raken aan wettelijk beschermd terrein en voeden verborgen proxy-discriminatie (Financial Times, 2025).

De EU AI Act verankert het governance-verhaal juridisch. Toepassingen in de arbeidsmarkt, kredietverlening, onderwijs en justitie worden expliciet als high-risk beschouwd; de bijbehorende verplichtingen, auditing, datakwaliteit, logging, documentatie, menselijk toezicht, worden van kracht in augustus 2026 en augustus 2027 (Europese Unie, 2024).

✦ KERNINZICHTEN

Algorithmische bias is geen marginaal technisch defect, maar een structureel risico dat rechtstreeks ingrijpt op menselijke kansen, rechten en waardigheid. Bias kan ontstaan in data, modellen, menselijke beoordeling, gekozen metrics en deploymentcontexten. Praktijkvoorbeelden zoals Amazon, COMPAS, verzekeringen en gezichtsherkenning tonen dat ogenschijnlijk objectieve systemen zeer oneerlijke effecten kunnen hebben. Automation bias maakt dat extra gevaarlijk, omdat mensen AI-outputs vaak te snel vertrouwen. Een verantwoorde aanpak vraagt daarom auditing, representatieve data, uitlegbaarheid, fairness-bewuste ontwerpkeuzes en sterke governance.

HOOFDSTUK 3

Privacy en surveillance

Wie kijkt mee, en in welke context

Controle over informatie en over het eigen leven

Wanneer mensen gevraagd worden of ze iets te verbergen hebben, is het antwoord doorgaans nee. Maar dat is de verkeerde vraag. Privacy gaat niet over geheimhouding. Ze gaat over controle: wie mag bepalen hoe informatie over jou wordt verzameld, gebruikt en gedeeld?

Die vraag krijgt een nieuwe urgentie in het tijdperk van artificiële intelligentie. AI-systemen verwerken gegevens op een schaal en snelheid die eerder ondenkbaar waren, en ze combineren databronnen op manieren die privacy ondermijnen zonder ook maar één geheim te onthullen.

Privacy als contextuele integriteit

Helen Nissenbaum (2004) formuleerde een definitie die verder gaat dan "niets te verbergen hebben". Zij introduceerde het begrip *contextuele integriteit*: informatie is gepast wanneer ze circuleert binnen de sociale context waarvoor ze bedoeld was. Medische gegevens zijn gepast bij een arts. Diezelfde gegevens zijn niet gepast bij een verzekeraar of werkgever, ook als de informatie zelf nooit geheim was.

Voor AI-systemen is dat onderscheid cruciaal. Systemen die data combineren uit sociale media, locatietracking, aankoopgedrag en online zoekgeschiedenissen produceren gedetailleerde profielen zonder ook maar één geheim gegeven vrij te geven. Nissenbaums kader maakt bovendien duidelijk waarom zogenaamd anonieme data in de praktijk zelden echt anoniem zijn: wie de woonplaats, werkplek en reispatronen van iemand kent, kan vaak een precies profiel opbouwen zonder ook maar één naam te kennen.

De economie van surveillance

Shoshana Zuboff beschrijft in *The Age of Surveillance Capitalism* (2019) hoe menselijke ervaring de grondstof is geworden van een nieuw economisch systeem. Bedrijven oogsten gedragsdata, zetten die om in voorspellingsmodellen en verhandelen die modellen, grotendeels buiten democratische controle en zonder expliciete toestemming van de betrokkenen.

Wat Zuboffs analyse bijzonder maakt, is de systeemlogica die ze blootlegt. Hoe meer data een model verwerkt, hoe nauwkeuriger de voorspellingen worden, en hoe sterker de economische prikkel om nóg meer data te verzamelen. Locatiegegevens, klikgedrag, communicatiepatronen en biometrische informatie zijn zo geen bijproduct van digitale diensten, ze zijn de kern van het businessmodel.

Spraakassistenten zoals Alexa of Google Assistant luisteren voortdurend mee om commando's te verwerken. Gebruiksvoorwaarden leggen ruime toestemmingen vast voor profilering en commerciële exploitatie, toestemmingen die de meeste gebruikers nooit bewust gelezen of begrepen hebben. De ethische vraag is of zulke toestemming vrij, geïnformeerd en betekenisvol kan zijn wanneer de alternatieven ontbreken.

Dezelfde logica functioneert in de arbeidscontext. Gedurende de coronapandemie verdubbelde het gebruik van software voor werknemersmonitoring nagenoeg overnight. Tools zoals Hubstaff, Teramind en de productiviteitsmodule van Microsoft Viva Insights meten toetsaanslagen, muisbewegingen, schermactiviteit en soms camerabeelden. De doelgroep was niet een kwetsbare minderheidsgroep, maar de doorsnee

kantoorwerknemer in sectoren als verzekeringen, bankwezen en openbaar bestuur. De asymmetrie is structureel: de werkgever beschikt over gedetailleerde gedragsdata; de werknemer heeft doorgaans onvoldoende zicht op wat precies gemeten wordt, hoe lang die data worden bewaard en hoe ze meewegen in beoordelingsgesprekken of contractbeslissingen.

In december 2025 berichtte AP News hoe de Amerikaanse grenspolitie via automatische nummerplaatherkenning en een voorspellend algoritme het reisgedrag van miljoenen bestuurders analyseerde, niet om concrete verdachten op te sporen maar om "verdachte" verplaatsingspatronen te markeren. Het is een sprekend voorbeeld van hoe surveillance op grote schaal werkt zonder dat betrokkenen daar weet van hebben.

In Europa biedt de slimme energiemeter een alledaagser voorbeeld. In nieuwe woningen in België en Nederland is een digitale meter met kwartieruuropnames wettelijk verplicht. Uit dat verbruikspatroon is nauwkeurig af te leiden wanneer iemand opstaat, wanneer hij thuis is, hoeveel mensen er in een huishouden wonen en welke apparaten in gebruik zijn. Formeel zijn die data geanonimiseerd, maar onderzoek toont dat individuele huishoudens op basis van verbruiksprofielen herleidbaar zijn, ook zonder naam of adres. De surveillance is hier niet opgelegd door een techbedrijf of een veiligheidsdienst: ze is ingebakken in een infrastructuur die de overheid als technische standaard heeft ingevoerd.

Gezichtsherkenning: techniek en ongelijkheid

Gezichtsherkenning brengt de spanning tussen veiligheid en privacy op haar scherpst. Systemen voor realtime biometrische identificatie maken het mogelijk om mensen fysiek te volgen en te koppelen aan databanken, in winkels, op straten, op luchthavens, vaak zonder dat de betrokkenen weten wanneer dat gebeurt of op welke grond.

Uit onderzoek van Wang et al. (2024) blijkt dat de nauwkeurigheid van gezichtsherkenningssystemen systematisch lager is bij mensen met een donkere huidskleur, bij vrouwen en bij ouderen. Dat is geen technische bijkomstigheid. Het weerspiegelt de samenstelling van de trainingsdata: systemen die voornamelijk getraind zijn op beelden van witte mannen presteren slechter op groepen die in die data ondervertegenwoordigd zijn. De technologie reproduceert zo de ongelijkheden die al in de samenleving bestaan.

De rechtsgevolgen zijn concreet. In februari 2026 werd een man in het Verenigd Koninkrijk gearresteerd voor een woninginbraak honderd kilometer van zijn huis, op basis van een foutieve gezichtsherkenningsidentificatie. De werkelijke dader was op de camerabeelden duidelijk jonger. Eind 2025 lobbyde de Britse politie om een soortgelijk systeem te behouden waarvan de hogere foutmarges voor vrouwen en niet-witte personen reeds bekend waren.

De EU AI Act (2024) geeft op dit punt een duidelijk wetgevend antwoord. Artikel 5 verbiedt realtime biometrische identificatie in publieke ruimten voor rechtshandhaving, met uitzondering van strikt afgebakende situaties zoals het zoeken naar vermiste slachtoffers of het afwenden van een onmiddellijke terroristische aanslag. Volledig verboden, zonder enige uitzondering, is het ongericht aanleggen van gezichtsherkenningsdatabanken via scraping van camerabeelden of internetfoto's.

Het chilling effect: hoe surveillance gedrag verandert

Een van de meest onderzochte gevolgen van surveillance is het zogenaamde chilling effect: de vaststelling dat bewustzijn van toezicht menselijk gedrag beïnvloedt, ook wanneer dat gedrag volledig legaal is.

Uit onderzoek van Büchi, Festic en Latzer (2022), gepubliceerd in *Big Data & Society*, blijkt dat meer dan de helft van de internetgebruikers aangaf minder vrij informatie op te zoeken of minder openlijk te communiceren zodra ze bewust werden van digitale monitoring. Penney (2022) analyseert in *Minnesota Law Review* het onderliggende mechanisme: wanneer mensen in een situatie van ambigu toezicht verkeren, censureren ze zichzelf preventief. Ze vermijden gedrag dat weliswaar legaal is, maar mogelijk als verdacht geïnterpreteerd zou kunnen worden.

De consequentie is politiek, niet alleen persoonlijk. Vrijheid van meningsuiting en van vereniging zijn grondrechten die in de praktijk pas werken wanneer mensen ze ook durven uitoefenen. Surveillance ondermijnt die praktische vrijheid zonder enige wet te overtreden. Mensen zoeken bepaalde dingen niet op. Ze sturen bepaalde berichten niet. Ze spreken zich niet uit.

Het Chinese sociale kredietsysteem toont waar die logica in haar meest uitgewerkte vorm toe leidt. Burgers worden beoordeeld op basis van financieel gedrag, online activiteit en camerabeelden. De score bepaalt hun toegang tot diensten, mobiliteit en maatschappelijke kansen. Surveillance functioneert daar niet langer als observatie maar als sturingsinstrument.

Regulering: mogelijkheden en grenzen

De GDPR legt principes vast zoals dataminimalisatie, doelbinding en beperkingen op geautomatiseerde besluitvorming. De EU AI Act voegt risicoclassificatie en specifieke verboden toe. Ze geven burgers en toezichthouders instrumenten die er voordien niet waren.

Toch toont de praktijk dat regelgeving structureel achterloopt op technologische ontwikkeling. In april 2026 herbekeken Amerikaanse steden contracten met Flock Safety, een bedrijf dat nummerplaatcamera's levert aan politiediensten, na bezorgdheid over wie toegang had tot de verzamelde gegevens en hoe lang die werden bewaard. De systemen waren al operationeel. Het debat begon pas daarna.

Een structurele beperking van Europese privacyregelgeving is de afhankelijkheid van niet-Europese cloudinfrastructuur. Belgische ziekenhuizen, overheidsdiensten en bedrijven die gebruikmaken van Microsoft 365, Google Workspace of Amazon Web Services slaan hun data op bij aanbieders die juridisch vallen onder de Amerikaanse CLOUD Act (2018). Die wet verleent Amerikaanse opsporingsdiensten het recht om data op te vragen, ongeacht de fysieke locatie van de servers. In 2025 erkende Microsoft voor de Franse Senaat dat het voor Europese klantdata geen absolute soevereiniteitsgarantie kan bieden. GDPR en de EU AI Act reguleren wat Europese actoren met data mogen doen; ze reguleren niet de juridische greep die een buitenlandse mogendheid op diezelfde data kan uitoefenen.

Er zijn ook technische benaderingen die de spanning tussen datagebruik en privacy deels kunnen verminderen. *Differential privacy* voegt wiskundig gecalibreerde ruis toe aan datasets, zodat individuele gegevenspunten niet herleid kunnen worden tot een specifieke persoon, terwijl statistische patronen op populatieniveau bruikbaar blijven. De techniek wordt onder meer toegepast door Apple en de Amerikaanse

volkstelling, en laat toe om inzichten uit gevoelige data te destilleren zonder die data ooit bloot te stellen. *Federated learning* werkt anders: het model wordt lokaal getraind op elk apparaat afzonderlijk, en alleen de gegeneraliseerde leeropbrengsten, niet de onderliggende persoonsgegevens, worden teruggestuurd naar een centrale server. Ruwe data verlaten het toestel nooit. *Versleutelde gegevensverwerking* maakt het in bepaalde gevallen mogelijk om berekeningen uit te voeren op versleutelde data zonder die eerst te ontsleutelen, zodat de inhoud ook voor de verwerker zelf onzichtbaar blijft.

De praktische toepasbaarheid op grote schaal blijft echter beperkt, en de fundamentelere beleidsvraag lossen deze technieken hoe dan ook niet op: welke vormen van verzameling en analyse zijn maatschappelijk aanvaardbaar, ongeacht wat technisch mogelijk is?

✦ KERNINZICHTEN

Privacy is een voorwaarde voor autonomie. Wie geen controle heeft over zijn gegevens, verliest gedeeltelijk ook de controle over zijn eigen leven. AI-systemen vergroten die kwetsbaarheid doordat ze op grote schaal gegevens combineren op manieren die afzonderlijk onschuldig lijken maar samen ingrijpende profielen opleveren. De gevolgen zijn ongelijk verdeeld: groepen die al ondervertegenwoordigd zijn in trainingsdata of minder institutionele bescherming genieten, worden disproportioneel getroffen. Surveillance heeft bovendien een preventief effect op gedrag, mensen censureren zichzelf wanneer ze zich bekeken weten, ook zonder concrete dreiging. Wetgeving biedt een kader, maar vereist actief toezicht en duidelijke normen over maatschappelijke aanvaardbaarheid om effectief te zijn.

HOOFDSTUK 4

Explainability

Beslissingen die zichzelf moeten kunnen verantwoorden

De uitleg die niemand vraagt

Een algoritme kan een krediet weigeren, een risicoscore toekennen aan een verdachte, een behandelingsadvies genereren. Het kan dat doen zonder een reden te geven die iemand kan begrijpen, betwisten of weerleggen. In menselijke besluitvorming is dat ondenkbaar: een rechter die vonnist zonder motivering, een arts die behandelt zonder diagnose, een overheid die weigert zonder grond. Voor AI-systemen is het standaard.

Dat is het vertrekpunt van explainability. De relevante vraag is niet of een systeem accuraat presteert, maar of de mensen die aan zijn beslissingen onderworpen zijn ze kunnen begrijpen, betwisten of weerleggen.

Transparantie, explainability en interpretability

Drie begrippen die vaak door elkaar worden gebruikt, betekenen drie verschillende dingen. *Transparantie* gaat over openheid rond systeemontwerp, gebruikte data en beslissingslogica. *Explainability* gaat over de redenen voor specifieke uitkomsten, begrijpelijk voor de ontvanger. *Interpretability* verwijst naar het menselijke vermogen om het gedrag van een model te volgen.

Finale Doshi-Velez en Been Kim stelden in een invloedrijk werkdocument (*Towards a Rigorous Science of Interpretable Machine Learning*, 2017) dat begrijpelijkheid contextafhankelijk is: uitleg die bruikbaar is voor een ingenieur, is dat niet noodzakelijk voor een patiënt of een rechter. Er bestaat geen universele maat voor goede uitleg; de geschiktheid van een verklaring hangt altijd af van de specifieke gebruiker, de specifieke context en de specifieke beslissing.

Zachary Lipton waarschuwde in "The Mythos of Model Interpretability" (*ACM Queue*, 2018) voor een fundamenteel onderscheid dat te vaak verdwijnt: het verschil tussen transparantie, waarbij men de interne werking van een model werkelijk kan volgen, en post-hoc rationalisaties, waarbij men achteraf verklaringen genereert die de werkelijke beslissingslogica niet weergeven. Een organisatie kan een overtuigende uitleg aanbieden die het model zelf niet raakt. Transparantie betekent dat het boek openligt. Explainability betekent dat het boek ook leesbaar is.

Black-boxmodellen en de asymmetrie van uitleg

Een centraal probleem is de *black box*: modellen, zoals diepe neurale netwerken of complexe ensemblemethoden, waarvan de interne werking zo hoog-dimensionaal is dat zelfs hun ontwikkelaars niet altijd kunnen reconstrueren hoe een specifieke beslissing tot stand kwam.

In alledaagse toepassingen is dat aanvaardbaar. In high-stakes contexten niet. Wanneer een systeem een krediet weigert, een recidiverisicoscore toekent of een behandelingsadvies genereert, is het ethisch moeilijk verdedigbaar dat niemand die beslissing kan reconstrueren. Systemen als COMPAS, het recidiverisicosysteem dat in hoofdstuk 2 aan bod komt, illustreren het probleem concreet: omdat de scoringslogica ondoorzichtig bleef, konden beklaagden de risicoscore niet effectief betwisten, ook niet wanneer die aantoonbaar beïnvloed was door ras.

Cynthia Rudin formuleerde in *Nature Machine Intelligence* (2019) een radicale conclusie: in high-stakes domeinen moeten ontwikkelaars stoppen met het achteraf uitleggen van black-boxmodellen en inherent interpreteerbare modellen bouwen. Haar argument is tweeledig. Ten eerste zijn post-hoc verklaringen per definitie een benadering van wat het model werkelijk doet, niet de werkelijkheid. Ten tweede is de veronderstelde accuraatheidswinst van complexe modellen ten opzichte van interpreteerbare alternatieven in de meeste high-stakes toepassingen empirisch niet aangetoond. Wie kiest voor een black box, aanvaardt niet alleen een uitlegbaarheidsnadeel, de premisse van de trade-off klopt al niet.

Die ongelijkheid in uitlegbaarheid is structureel: ze volgt de lijnen van kwetsbaarheid. Wie institutionele macht heeft, juridische bijstand, technische kennis, middelen om te procederen, kan een beslissing betwisten. Wie die macht niet heeft, kan dat niet. Virginia Eubanks toonde in *Automating Inequality* (2018) hoe dit mechanisme werkt voor publieke dienstverlening: mensen in armoede worden blootgesteld aan geautomatiseerde beoordeling terwijl hun de instrumenten ontbreken om die beoordeling te contesteren.

Technieken voor uitlegbaarheid

De praktijk van explainable AI biedt instrumenten om black-boxmodellen beheersbaar te maken. Marco Tulio Ribeiro en collega's stelden in 2016 LIME voor, *Local Interpretable Model-Agnostic Explanations*. Voor één specifieke voorspelling bouwt LIME een lokaal, eenvoudig model dat het gedrag van het complexe model in de nabijheid van dat datapunt benadert. Zo wordt zichtbaar welke kenmerken voor die ene beslissing doorslaggevend waren.

Scott Lundberg en Su-In Lee introduceerden in 2017 SHAP, *SHapley Additive exPlanations*, gebaseerd op de Shapley-waarden uit de speltheorie. SHAP verdeelt de bijdrage van elk kenmerk aan een voorspelling op een wiskundig consistente manier en biedt een coherenter kader dan LIME voor zowel lokale als globale interpretatie.

Counterfactuals tonen welk element had moeten veranderen om een andere uitkomst te krijgen: "Als uw inkomen 10% hoger was geweest, had u de lening wel gekregen." *Saliency maps* visualiseren in neurale netwerken welke zones of kenmerken de meeste aandacht kregen, gangbaar in beeldherkenning en medische beeldverwerking.

Lipton (2018) wees erop dat post-hoc verklaringen het publiek een gevoel van begrip kunnen geven zonder dat dit begrip ook echt wordt overgedragen. LIME en SHAP zijn nuttige instrumenten voor documentatie en diagnose; ze zijn geen vervanging voor modellen die van zichzelf begrijpelijk zijn.

Datadocumentatie en model cards

Uitlegbaarheid begint vóór het model bestaat: bij de data. Timnit Gebru en collega's introduceerden in 2021 *datasheets for datasets*, een gestandaardiseerd documentatieformaat dat beschrijft hoe een dataset werd samengesteld, met welk doel, voor welke groepen ze representatief is en welke beperkingen ze heeft. Dergelijke documentatie maakt het mogelijk om systematische problemen vroeg op te sporen, vóór een model getraind en ingezet wordt.

Margaret Mitchell en collega's stelden in 2019 *model cards* voor: gestandaardiseerde documenten die beschrijven waarvoor een model bedoeld is, hoe het presteert voor verschillende subgroepen, welke beperkingen het heeft en welk gebruik ervan ongeschikt of gevaarlijk is. Samen met datasheets bieden model cards een transparantie-infrastructuur die verder gaat dan technische specificaties en ook ethische context meeneemt.

Volledige openheid is niet altijd mogelijk: in medische contexten kan men niet zomaar alle trainingsdata publiek maken. Ook daar is een balans nodig tussen transparantie, privacy en praktische haalbaarheid, maar die afweging moet expliciet gemaakt worden, niet stilzwijgend vermeden.

Juridische verplichtingen

Explainability is in Europa geen ethische voorkeur maar een afdwingbare juridische verplichting. Artikel 22 van de GDPR verbiedt geautomatiseerde beslissingen met rechtsgevolgen of vergelijkbare significante gevolgen voor individuen, tenzij aan strikte voorwaarden is voldaan. De betrokkene heeft recht op menselijke tussenkomst, het uiten van zijn standpunt en het betwisten van de beslissing. De Artikelen 13 tot 15 GDPR verplichten bovendien tot "zinvolle informatie over de gehanteerde logica."

Het Hof van Justitie van de Europese Unie preciseerde op 27 februari 2025 (zaak C-203/22) de reikwijdte van die verplichting: verwerkingsverantwoordelijken kunnen niet volstaan met de "loutere mededeling van een complexe wiskundige formule, zoals een algoritme, of door de gedetailleerde beschrijving van alle stappen in de geautomatiseerde besluitvorming." Ze moeten de werkelijk toegepaste procedure en de reële principes uitleggen die op de gegevens van de betrokkene werden toegepast.

De EU AI Act versterkt die redenering. Artikel 13 legt aanbieders van hoog-risico AI-systemen de plicht op hun systemen zodanig te ontwerpen dat deployers de output kunnen interpreteren en gepast kunnen gebruiken. Providers moeten duidelijke instructies leveren over mogelijkheden, beperkingen en potentiële risico's. De verplichtingen voor hoog-risico systemen, waaronder toepassingen in kredietverlening, aanwerving, onderwijs en strafrecht, worden van kracht in augustus 2026.

✦ KERNINZICHTEN

Explainability verwijst naar het vermogen van een AI-systeem om begrijpelijke redenen te geven voor beslissingen, afgestemd op de gebruiker en de context. Transparantie, explainability en interpretability zijn verwante maar onderscheiden begrippen: transparantie gaat over openheid van het systeem, explainability over verstaanbare uitleg van specifieke beslissingen, interpretability over het menselijk vermogen om modelgedrag te volgen. Black-boxmodellen vormen een ethisch risico in high-stakes domeinen, niet omdat uitleg technisch onmogelijk is, maar omdat post-hoc verklaringen per definitie benaderingen zijn van wat het model werkelijk doet. Rudin (2019) toont dat inherent interpreteerbare modellen in veel high-stakes contexten vergelijkbare prestaties halen als complexe alternatieven; de veronderstelde trade-off tussen accuraatheid en uitlegbaarheid is grotendeels een aanname. De last van onuitlegbaarheid valt niet gelijk: wie al kwetsbaar is, mist ook de middelen om een ondoorzichtige beslissing te betwisten. Het Hof van Justitie van de EU (zaak C-203/22, 2025) en de EU AI Act (Artikel 13) maken explainability tot een afdwingbare verplichting, niet alleen een technische wens.

HOOFDSTUK 5

AI Governance

Wie bestuurt de besturing

Inleiding

Een systeem dat niemands verantwoordelijkheid is, is niemands probleem. Dat is de kern van de governancevraag: zodra AI-systemen beslissingen ondersteunen of automatiseren die rechten, kansen of veiligheid van mensen raken, stelt zich niet alleen de vraag wat ze doen, maar ook wie ervoor instaat dat ze het goed doen, en wie aansprakelijk is wanneer dat niet lukt.

Governance gaat over die verdelingsvraag. Wie beslist welke systemen toegestaan zijn? Wie controleert of ze werken zoals beloofd? Wie draagt de gevolgen wanneer ze falen? Die vragen zijn evenzeer politiek als technisch, en ze kunnen niet worden overgelaten aan de markt of aan de technische teams die de systemen ontwikkelen.

Wat AI governance betekent

AI governance kan worden begrepen als het geheel van regels, principes, instellingen en praktijken dat bepaalt hoe AI ontwikkeld en ingezet mag worden. Allan Dafoe omschreef in zijn invloedrijke onderzoeksagenda (2018, Centre for the Governance of AI, Oxford) AI governance als het geheel van "mechanisms and processes that influence whether AI research and applications go well". Die definitie benadrukt meteen dat governance verder gaat dan regels schrijven: het gaat over wie invloed heeft op de richting van AI, via welke kanalen en met welke middelen. Het onderwerp draait dus niet alleen om overheidstoezicht, maar ook om de interne keuzes van bedrijven, de rol van ethische commissies, de verwachtingen van burgers en de internationale afspraken die landen proberen te maken.

Governance is daarmee breder dan regulering alleen. Regulering verwijst naar bindende regels en wettelijke kaders. Governance omvat ook soft law, vrijwillige richtlijnen, industriestandaarden, auditing, documentatie, training en interne besluitvorming, instrumenten die nodig zijn juist omdat AI sneller evolueert dan klassieke wetgeving kan volgen.

Kernprincipes van AI governance

Uit een systematische vergelijking van 84 ethische AI-richtlijnen van overheden, bedrijven, onderzoekscentra en internationale organisaties, uitgevoerd door Anna Jobin, Marcello Ienca en Effy Vayena en gepubliceerd in Nature Machine Intelligence (2019), blijkt dat er een globale convergentie bestaat rond vijf kernprincipes: transparantie, rechtvaardigheid en fairness, niet-schadelijkheid, verantwoordelijkheid en privacy. Die convergentie is betekenisvol: ze suggereert dat er een breed gedeelde morele intuïtie bestaat over wat van AI-systemen verwacht mag worden. Jobin en collega's stelden echter ook vast dat achter die gedeelde begrippen grote inhoudelijke divergentie schuilgaat. Wat precies "transparantie" of "fairness" inhoudt, hoe die principes gerelateerd zijn, voor wie ze gelden en hoe ze geoperationaliseerd worden, verschilt aanzienlijk per document, regio en sector.

Transparantie impliceert dat AI-systemen voldoende begrijpelijk zijn om hun werking en beslissingen te verantwoorden, niet dat elk technisch detail altijd voor iedereen zichtbaar moet zijn, maar dat betrokkenen en toezichthouders niet volledig afhankelijk mogen zijn van een black box. Accountability vereist dat wanneer een AI-systeem schade veroorzaakt, discrimineert of onveilig blijkt, duidelijk is wie verantwoor-

delijk is: ontwikkelaars, deployers, gebruikers, management of toezichthouders. Zonder die toewijzing blijft ethische controle zwak en worden fouten niet gecorrigeerd. Fairness, veiligheid en privacy vullen het kader aan: een systeem mag geen bestaande ongelijkheden versterken zonder dat dit zichtbaar of betwistbaar is, moet robuust zijn tegen fouten en manipulatie, en persoonsgegevens en privacyrechten respecteren. Luciano Floridi en collega's voegden in het AI4People-rapport (Minds and Machines, 2018) een vijfde principe toe: explicability, AI moet niet alleen goed en eerlijk zijn, maar ook begrijpelijk genoeg om op goede gronden te kunnen worden vertrouwd of betwist.

Soft law, hard law en voorzichtigheid

Soft law, niet-bindende richtlijnen, codes of conduct en ethische aanbevelingen, kan snel worden ontwikkeld en aangepast, maar mist afdwingbaarheid. Hard law, juridisch bindende regels zoals GDPR of sectorspecifieke verplichtingen, biedt rechtszekerheid, maar evolueert trager. AI governance beweegt zich tussen beide. Wettelijke grenzen zijn nodig; vrijwillige kaders geven sneller richting aan een technologie die sneller verandert dan wetgeving.

Een risk-based approach sluit daarbij aan. Niet elk AI-systeem vraagt dezelfde mate van controle. Hoe groter het risico voor rechten, veiligheid of maatschappelijke schade, hoe zwaarder de governance en het toezicht moeten zijn. Het precautionary principle versterkt dat: wanneer risico's nog niet volledig begrepen worden, is voorzichtigheid gerechtvaardigd, vooraf grenzen trekken in gevoelige contexten is te verkiezen boven wachten tot de schade zichtbaar is.

De kritiek op de soft law-aanpak is fundamenteel. Thilo Hagendorff analyseerde in Minds and Machines (2020) 22 AI-ethische richtlijnen en stelde vast dat de meeste principes te vaag, te abstract en te weinig operationeel zijn om echte gedragsverandering te bewerkstelligen. De richtlijnen focussen doorgaans op technische begrippen als fairness en transparantie, maar negeren structurele machtsvragen: wie controleert AI-bedrijven, wie beschermt klokkenluiders, wie heeft toegang tot audits? Die omissies zijn geen toeval, aldus Hagendorff: ze weerspiegelen de invloed van de techsector op het ethische debat zelf, ethics washing (Metzinger): ethische kaders die dienen als vervanging voor regulering, in plaats van als aanvulling erop.

Governance in de gezondheidszorg

In de gezondheidszorg is de foutmarge het kleinst. AI-systemen worden er ingezet voor diagnose, behandelingsplanning, triage en predictieve analyses. Wanneer een model een aandoening verkeerd inschat, een behandeling verkeerd prioriteert of een groep patiënten systematisch anders beoordeelt, heeft dat directe gevolgen voor gezondheid en leven.

Governance in deze sector is dienovereenkomstig streng. In de VS classificeert de FDA AI-gebaseerde medische hulpmiddelen op basis van bestaande wetgeving voor medische software en vereist premarketbeoordeling voor systemen met hoog risico. De EU AI Act beschouwt AI in de gezondheidszorg als high-risk: aanbieders zijn verplicht tot conformiteitsbeoordeling, technische documentatie, transparantie en menselijk toezicht vóór marktintroductie. De verplichtingen zijn niet bedoeld om innovatie te blokkeren, maar om te voorkomen dat systemen met onbekende foutenmarges worden genormaliseerd voordat de schade zichtbaar wordt.

Governance in de financiële sector

In kredietverlening, fraudedetectie en geautomatiseerde financiële analyses neemt AI beslissingen met directe en materiële gevolgen: wie een lening krijgt, wiens betaling wordt geblokkeerd, wie als risicovol wordt gecategoriseerd. Modellen die getraind zijn op historische data reproduceren ook historische ongelijkheden. Wanneer bovendien te veel instellingen op vergelijkbare modellen steunen, kunnen systeemrisico's versterken en de markt breder destabiliseren.

De EU AI Act classificeert AI-systemen voor kredietbeoordeling van particulieren als high-risk en verplicht aanbieders tot transparantie over gebruikte data, logging van beslissingen en menselijk toezicht bij significante individuele besluiten. Kredietbeslissingen zijn daarmee juridisch betwistbaar terrein geworden, ook wanneer ze geautomatiseerd tot stand komen.

Governance bij autonome voertuigen

Autonome voertuigen leggen een specifiek governanceprobleem bloot: aansprakelijkheid bij realtime besluitvorming. Een systeem dat wegen interpreteert, obstakels detecteert en in fracties van seconden handelt, neemt in noodsituaties beslissingen die vroeger enkel mensen namen. Wanneer het misloopt, is onmiddellijk onduidelijk bij wie de verantwoordelijkheid ligt, de fabrikant, de operator, de eigenaar of de overheid die het systeem toeliet.

Safety standards, testverplichtingen, aansprakelijkheidskaders en beperkingen op dataverzameling vormen hier het governancekader. Governance richt zich daarbij op preventieve drempels vóór brede uitrol. De ethische vragen die dat oproept over algoritmische besluitvorming in noodsituaties zijn niet louter filosofisch: ze bepalen mee welke systemen wettelijk toegelaten mogen worden en onder welke voorwaarden.

Sociale media als governanceprobleem

Sociale mediaplatformen zijn een van de meest complexe governancecasussen. AI beheert er contentmoderatie, aanbevelingsalgoritmen en gerichte advertenties, op een schaal van miljarden gebruikers, in realtime, zonder transparantie over de onderliggende criteria. Privacyproblemen, ongelijke zichtbaarheid, versterking van polarisering en de verspreiding van desinformatie zijn daarbij geen bijverschijnselen maar structurele eigenschappen van de systemen.

Bestaande instrumenten proberen een kader te scheppen. De GDPR regelt gegevensverwerking. De Europese Digital Services Act (DSA, 2022) verplicht grote platformen tot transparantie over aanbevelingssystemen en risicobeoordelingen. De handhavingsvraag blijft echter open: wie controleert systemen die sneller evolueren dan wetgeving, op een schaal die elk toezichthoudersbureau overbelast?

Waarom regulering noodzakelijk is

Zelfregulering door de technologiesector heeft een gedocumenteerd tekortschietend spoor: ethische codes die niet worden nageleefd, richtlijnen die niet worden afgedwongen, auditresultaten die niet worden gepubliceerd. Wanneer de commerciële prikkel tot snelle uitrol botst met de vereiste voor grondige veiligheidstesting, wint doorgaans de marktlogica.

Regulering verschuift die prikkelstructuur. Wanneer organisaties moeten documenteren, testen, auditen en uitleggen, en dat aantoonbaar moeten maken aan een toezichthouder, stijgt de kans dat problemen vroeger worden opgespoord. Governance wordt daarmee een manier om verantwoordelijkheid proactief te organiseren, in plaats van reactief schade te beheren nadat systemen al op grote schaal zijn ingezet.

Belangrijke regelgevende kaders

De EU AI Act, formeel Verordening (EU) 2024/1689, is het eerste bindende wettelijke kader voor AI ter wereld en trad in werking in augustus 2024. Het kader vertrekt vanuit een risicobenadering met vier niveaus. Systemen met onaanvaardbaar risico (unacceptable risk) zijn verboden: dat omvat onder meer sociale scoring door overheden, realtime biometrische identificatie in publieke ruimte voor wetshandhaving (met beperkte uitzonderingen), en manipulatieve technieken die het vrije oordeel van personen ondermijnen. Hoog-risico systemen (high risk) in domeinen zoals gezondheidszorg, justitie, grensbewaking, onderwijs en arbeid zijn toegestaan, maar onderworpen aan strenge verplichtingen: conformiteitsbeoordeling, technische documentatie, transparantieregistratie, menselijk toezicht en post-market monitoring. Beperkt risico (limited risk) omvat systemen zoals chatbots, die enkel transparantieverplichtingen dragen. Minimaal risico (minimal risk), de meeste AI-toepassingen, valt buiten directe verplichtingen.

Naast de EU AI Act blijft GDPR een fundamenteel kader, ook al is die niet specifiek voor AI geschreven. Veel AI-systemen verwerken persoonlijke data, waardoor principes zoals toestemming, dataminimalisatie, transparantie en verantwoordelijkheid rechtstreeks relevant blijven. Een derde referentiepunt is het NIST AI Risk Management Framework (2023), ontwikkeld door het Amerikaanse National Institute of Standards and Technology. Dat vrijwillige kader organiseert AI-risicobeheer rond vier functies: Govern, Map, Measure en Manage. Het is minder prescriptief dan de EU AI Act, maar biedt organisaties een praktisch en sectoronafhankelijk instrument. GDPR, de EU AI Act en het NIST RMF weerspiegelen elk op een eigen manier het zoeken naar een evenwicht tussen innovatie, rechtsbescherming en accountability, waarbij de Europese aanpak bindend en grondrechtengebaseerd is, de Amerikaanse vrijwilliger en sectoraal.

Internationale AI governance

AI governance stopt niet aan de grens. Tussen de drie grote machtsblokken, de Europese Unie, de Verenigde Staten en China, bestaan fundamenteel verschillende benaderingen. De EU zet in op grondrechten, risicoregulering en juridische afdwingbaarheid. De VS heeft tot voor kort primair ingezet op sectorale zelfregulering en vrijwillige kaders, hoewel executive orders en sectorale wetgeving toenemen. China combineert een centralistische staatsaanpak met sectorale AI-regels, maar geeft daarin de prioriteit aan staatscontrole en nationale veiligheidsdoelstellingen boven individuele rechten.

In dat gefragmenteerde landschap spelen twee internationale kaders een coördinerende rol. De OECD AI Principles (2019, herzien 2024) zijn het eerste intergouvernementele normenkader voor AI, ondertekend door 42 landen. Ze omvatten vijf waarden-gebaseerde principes, inclusief transparantie, accountability, veiligheid en inclusieve groei, en vijf aanbevelingen aan beleidsmakers. De UNESCO-aanbeveling over AI-ethiek (2021) is het eerste mondiale normdocument dat door alle 193 UNESCO-lidstaten werd

aangenomen. Het richt zich nadrukkelijk op ongelijkheid, genderdimensies, milieu-impact en de rechten van gemarginaliseerde groepen, en pleit voor een moratorium op AI-toepassingen die mensenrechten in gevaar brengen zolang er geen adequate beschermingsmechanismen zijn.

Die internationale kaders zijn juridisch niet bindend. Ze bieden wel een gedeelde referentietaal die diplomatieke samenwerking en minimale normenconvergentie mogelijk maakt.

Problemen in globale regulering

Een terugkerend thema is dat globale regulering structureel worstelt met versnippering. Verschillende landen gebruiken andere definities, andere prioriteiten en andere handhavingsmechanismen. Daardoor ontstaan regulatory gaps en accountability gaps. Een AI-systeem kan in meerdere landen actief zijn, maar nergens volledig helder onder een coherent regime vallen.

Geopolitieke spanningen maken dat nog moeilijker. De Verenigde Staten en China volgen bijvoorbeeld niet dezelfde benadering, wat wereldwijde coördinatie bemoeilijkt. Daar komt bij dat AI sneller evolueert dan wetgeving. Tegen dat een kader is uitgewerkt, kan de technologie al weer verschoven zijn. Regulatory capture vormt een bijkomend structureel risico: de directe invloed van private belangen op het reguleringsproces dat hen verondersteld wordt te controleren.

AI governance in ontwikkelingslanden

Een belangrijke nuance is dat AI governance niet overal vanuit dezelfde uitgangspositie vertrekt. In ontwikkelingslanden zijn er vaak minder middelen, minder gespecialiseerde expertise en soms zwakkere institutionele capaciteit om complexe AI-systemen te reguleren. Daardoor dreigt een ongelijke wereldorde waarin sommige regio's technologie produceren en normeren, terwijl andere regio's vooral de risico's ondergaan.

Governance-modellen die zijn ontworpen voor hoogindustrialiseerde economieën met sterke rechtsstaten zijn bovendien niet automatisch overdraagbaar. Rechtenbescherming, inclusie en maatschappelijke relevantie vragen andere accenten naargelang infrastructuur, economie en politieke situatie. Dat maakt globale AI governance structureel complex, en maakt internationale dialoog over minimumnormen des te urgenter.

Cross-border data governance en Europese digitale soevereiniteit

De governancevraag wordt nog complexer zodra data grenzen overschrijden. Een bijzonder acuut spanningspunt is de botsing tussen de GDPR en de Amerikaanse CLOUD Act (Clarifying Lawful Overseas Use of Data Act, 2018). De CLOUD Act verplicht Amerikaanse bedrijven en clouddiensten om op rechtmatig verzoek van Amerikaanse autoriteiten toegang te verlenen tot data die ze bezitten of beheren, ook wanneer die data buiten de VS opgeslagen is. Dat botst direct met de GDPR, die doorgifte van persoonsgegevens aan derde landen aan strikte voorwaarden onderwerpt. Europese organisaties die gebruikmaken

van Amerikaanse cloudinfrastructuur (zoals Microsoft Azure, Amazon Web Services of Google Cloud) bevinden zich daardoor in een rechtsonzekere positie: ze kunnen tegelijkertijd verplicht zijn om data te beschermen onder GDPR en verplicht zijn om toegang te verlenen onder de CLOUD Act.

Die spanning voedt een bredere politieke discussie over Europese digitale soevereiniteit: de ambitie om de afhankelijkheid van de Europese digitale infrastructuur van buitenlandse, en in het bijzonder Amerikaanse, technologiebedrijven te verminderen. Initiatieven zoals GAIA-X, de Europese dataspaces en de Data Governance Act (2022) zijn concrete pogingen om een Europees dataecosysteem op te bouwen dat zowel technologisch competitief als juridisch autonoom is. Dat streven heeft ook directe implicaties voor AI governance: wie de data controleert, controleert ook de modellen die erop getraind worden.

Vergelijkbare spanningen doen zich voor met China's Personal Information Protection Law (PIPL, 2021), die gegevensdoorgifte naar buiten China reguleert vanuit een geheel andere politieke logica. Daarom duiken voorstellen op rond internationale interoperabiliteitsafspraken voor datastromen, maar vooralsnog blijft AI governance op dit vlak gefragmenteerd langs nationale en geopolitieke lijnen.

Case study: Clearview AI

Clearview AI verzamelde zonder toestemming meer dan dertig miljard gezichtsafbeeldingen van het publieke internet en bood die databank aan als opsporingsdienst voor politiediensten wereldwijd. In meerdere Europese landen, waaronder België, Italië en Griekenland, legden toezichthouders miljoenenboetes op wegens schending van de GDPR. In 2022 kondigde het bedrijf aan zijn diensten niet langer aan te bieden in Europa.

De casus maakt een governancepatroon zichtbaar dat vaker voorkomt: een technologie wordt grootschalig ingezet, de juridische kaders blijken onvoldoende of niet gehandhaafd, en pas na maatschappelijke druk en klachten van toezichthouders volgt een reactie. De EU AI Act poogt dat patroon te doorbreken door biometrische identificatiedatabanken die via scraping zijn aangelegd expliciet te verbieden, zonder enige uitzondering. Of dat verbod effectief wordt afgedwongen, zal afhangen van de capaciteit en politieke wil van nationale toezichthouders.

Governance binnen organisaties

Wetgeving bepaalt de buitengrens van wat toegestaan is. Wat er daarbinnen gebeurt, hangt af van de interne structuren van organisaties die AI ontwikkelen of inzetten. Die structuren zijn doorgaans zwakker dan de publieke governance-architectuur suggereert.

Effectieve interne governance veronderstelt drie dingen: duidelijke verantwoordelijkheidsverdeling, wie beslist over de inzet van een systeem, en wie controleert of dat besluit terecht was, multidisciplinaire input vóór de inzetbeslissing, en transparante rapportering wanneer systemen niet functioneren zoals verwacht. De EU AI Act verplicht high-risk aanbieders tot precies die structuren: conformiteitsbeoordeling, technische documentatie en menselijk toezicht zijn geen vrijwillige goede praktijken maar wettelijke verplichtingen.

Interne processen en compliance

Data governance, weten welke data verzameld worden, hoe lang ze bewaard worden en wie er toegang toe heeft, is een fundament van interne AI governance. Zonder die basiscontrole zijn bias auditing en accountability niet operationaliseerbaar. Bias auditing test of systemen ongelijke uitkomsten produceren voor verschillende groepen, zowel voor als na deployment. Algorithmic transparency vereist dat de beslissingslogica van een systeem intern gedocumenteerd en aantoonbaar is, voor de verantwoordelijken en voor toezichthouders.

Compliance is geen eindpunt. De EU AI Act, GDPR en sectorale verplichtingen evolueren mee met de technologie en met politieke prioriteiten. Organisaties die governance behandelen als een eenmalig conformiteitsproces riskeren dat systemen die vandaag voldoen, morgen in een grijs gebied terechtkomen.

Monitoring, training en tools

Governance eindigt niet bij de inzetbeslissing. Systemen die na deployment niet worden opgevolgd, kunnen ongemerkt driften, in prestatie, in bias, in afstemming met een evoluerende regelgeving. Post-deployment monitoring omvat technische prestatiemeting, bias-tracking over tijd en periodieke herziening van de inzetcontext.

AI literacy bij medewerkers is een onderschatte component. Governanceprocessen die alleen op papier bestaan, ethische richtlijnen die niemand gelezen heeft, auditchecklists die formeel worden afgevinkt, zijn geen governance. Mensen moeten de risico's begrijpen van de systemen waarmee ze werken, en weten hoe ze bezorgdheden intern kunnen melden.

Technische tools zoals Fairlearn, SHAP en LIME ondersteunen biasdetectie en explainability. Ze zijn hulpmiddelen, geen vervanging voor de institutionele structuur die governance draagt.

Meten van succes

Governance die niet meetbaar is, blijft symbolisch. Compliance rates, auditfrequentie, biasreductie over tijd en het aantal gevallen waarbij menselijk toezicht daadwerkelijk werd ingeschakeld, dat zijn indicatoren die concrete informatie geven over of een systeem in de praktijk naar behoren functioneert. Zonder zulke indicatoren bestaat het risico dat organisaties governance verwisselen met het hebben van een ethisch charter.

Goede AI governance is geen perfecte controle. Ze is een structurele inspanning om risico's vroegtijdig te detecteren, verantwoordelijkheid niet te verliezen aan de complexiteit van systemen, en bij te sturen wanneer systemen niet functioneren zoals bedoeld.

✦ KERNINZICHTEN

AI governance omvat het geheel van regels, principes, instellingen en praktijken dat bepaalt hoe AI wordt ontwikkeld en ingezet, breder dan regulering alleen, en altijd verweven met politieke keuzes over wie beslist en voor wie. Een mondiale convergentie bestaat rond vijf kernprincipes (Jobin et al., 2019): transparantie, fairness, niet-schadelijkheid, verantwoordelijkheid en privacy; achter die gedeelde taal gaat echter grote inhoudelijke divergentie schuil over interpretatie en implementatie. Soft law-kaders bieden snelheid en flexibiliteit, maar riskeren ethics washing wanneer ze dienen als vervanging voor bindende regels; hard law zoals de EU AI Act biedt rechtszekerheid en vraagt politieke wil en afdwingingsmechanismen (Hagendorff, 2020). De EU AI Act (2024) is het eerste bindende wettelijke kader voor AI ter wereld en hanteert een risicogebaseerde benadering met vier categorieën, van verboden toepassingen tot minimaal risico. Internationaal ontbreekt coherentie: de VS, EU en China hanteren fundamenteel verschillende benaderingen, terwijl OECD AI Principles (2019) en UNESCO-aanbeveling (2021) proberen een gemeenschappelijke minimumnorm te bieden. De spanning tussen GDPR en de US CLOUD Act maakt Europese digitale soevereiniteit een urgente governance-kwestie. Effectieve governance vraagt interne structuren, auditing, training en meetbare opvolging, principes op papier volstaan niet.

HOOFDSTUK 6

GDPR en AI-Ethiek

Het juridische fundament onder ethisch AI-gebruik

Inleiding

Juridische naleving en ethische verantwoordelijkheid zijn niet hetzelfde. Een AI-systeem kan aan alle wettelijke vereisten voldoen en toch fundamentele waarden schenden: menselijke waardigheid reduceren tot een score, autonomie uithollen door ondoorzichtige automatisering, of groepen systematisch anders behandelen zonder dat iemand daarvoor juridisch aansprakelijk gesteld kan worden.

De GDPR legt vast wat organisaties minimaal moeten naleven bij de verwerking van persoonsgegevens. AI-ethiek voegt daaraan een normatieve laag toe: ze vraagt niet alleen of een systeem de regels volgt, maar welke waarden het dient, welke schade het kan veroorzaken, en hoe menselijke autonomie en waardigheid structureel kunnen worden beschermd.

GDPR als basis voor verantwoord AI-gebruik

GDPR vormt het juridische fundament voor een groot deel van ethisch AI-gebruik in Europa. De verordening dwingt organisaties om zorgvuldig om te gaan met persoonsgegevens, transparant te zijn over verwerking en rechten van betrokkenen te respecteren.

In de context van AI is dat bijzonder relevant, omdat veel AI-systemen steunen op grote hoeveelheden data. Zodra die data herleidbaar is tot personen, of gebruikt wordt om personen te beoordelen, profileren of classificeren, ontstaat een directe koppeling met gegevensbescherming.

GDPR is dus geen louter administratief kader. Het legt de basis voor privacybescherming, beperkt willekeurige dataverwerking en creëert waarborgen tegen ondoorzichtige geautomatiseerde besluitvorming. Voor AI-systemen waarbij meerdere actoren betrokken zijn bij ontwerp, training en inzet is die verdeling van rollen en verantwoordelijkheden bijzonder relevant: elk van die actoren draagt eigen juridische verplichtingen.

Belangrijke GDPR-principes voor AI

Kernprincipes zoals purpose limitation, accuracy en data minimization zijn in AI-context bijzonder belangrijk, omdat modellen anders de neiging hebben om zoveel mogelijk data te verzamelen, te hergebruiken en op te slaan.

Purpose limitation betekent dat persoonsgegevens alleen voor een duidelijk omschreven doel mogen worden gebruikt. Data minimization houdt in dat men niet meer data verzamelt dan nodig is. Accuracy verplicht organisaties om redelijke inspanningen te doen om met correcte gegevens te werken.

Samen zorgen die principes ervoor dat AI-systemen niet eindeloos gevoed worden met overbodige of slecht afgebakende persoonsgegevens. Dat is zowel juridisch verplicht als ethisch zinvol: het verkleint de kans op disproportionele monitoring of foutieve conclusies.

GDPR en geautomatiseerde besluitvorming

Een van de meest relevante onderdelen voor AI is het GDPR-kader rond automated decision-making en profiling, met onder meer Article 22 en Recital 71.

De kern van die bescherming is dat mensen rechten hebben wanneer belangrijke beslissingen volledig of grotendeels geautomatiseerd worden genomen. Organisaties moeten informeren over het gebruik van een dergelijk systeem, de onderliggende logica en het doel van de verwerking. Menselijke tussenkomst blijft daarin een cruciale waarborg.

Recruitment AI, credit scoring en predictive policing maken dit bijzonder concreet. Juist in zulke domeinen is uitleg noodzakelijk, omdat de uitkomst een directe invloed heeft op kansen, rechten en maatschappelijke positie.

In de academische literatuur is de precieze reikwijdte van dit recht echter betwist. Sandra Wachter, Brent Mittelstadt en Luciano Floridi (2017) betoogden in *International Data Privacy Law* dat Artikel 22 en Recital 71 van de GDPR strikt genomen geen volledig recht op uitleg van een concrete beslissing verlenen, maar eerder een recht om geïnformeerd te worden over de logica en de voorziene gevolgen van het systeem. Ze noemden dit een 'right to be informed' in plaats van een echte uitlegverplichting. Die nuance is juridisch relevant: aanvullende instrumenten, zoals de EU AI Act, zijn nodig om effectieve transparantierechten te realiseren.

AI-ethiek als laag boven compliance

Juridische naleving alleen volstaat niet. Een systeem kan formeel aan regels voldoen en toch ethisch problematisch blijven. Daarom is AI-ethiek een tweede, noodzakelijke laag.

Die ethische laag draait om vragen zoals: respecteert een systeem menselijke autonomie? Vermijdt het onnodige schade? Is het fair voor verschillende groepen? Worden mensen niet gereduceerd tot datapunten of scores zonder dat zij nog invloed hebben op de uitkomst?

GDPR legt de ondergrens vast; AI-ethiek helpt hogere standaarden van mensgericht ontwerp en gebruik te formuleren.

Privacy als morele en juridische plicht

Privacy is meer dan een juridisch recht. Ze is ook een morele imperatief die menselijke waardigheid, individuele vrijheid en autonomie beschermt.

Informed consent, beperkte dataverzameling en respect voor gebruikerscontrole zijn zowel juridische verplichtingen als ethische eisen. Organisaties die meer data verzamelen dan nodig of onduidelijk blijven over gebruiksdoelen, schenden een regel, maar ondermijnen daarmee ook het vertrouwen en de morele legitimiteit van hun systeem.

Het recht op uitleg

Het recht op uitleg verdient hierbij bijzondere aandacht. Gebruikers moeten betekenisvolle verklaringen kunnen krijgen over AI-beslissingen die hen treffen. Dat ondersteunt fairness, contestability en transparantie.

Financiële en medische aanbevelingen maken dit concreet. Als een AI-systeem een krediet weigert of een behandelvoorstel doet, volstaat een kale uitkomst niet. De betrokkene moet in zekere mate kunnen begrijpen waarom die beslissing of aanbeveling tot stand kwam en welke factoren meespeelden.

Lilian Edwards en Michael Veale (2017) wezen erop in de *Duke Law & Technology Review* dat een individueel uitlegrecht alleen niet volstaat als governance-instrument. Zij pleitten voor een bredere aanpak die ook Data Protection Impact Assessments, rechterlijke toetsing en modelregisters omvat. Transparantie moet structureel worden verankerd, niet alleen geregeld via ad-hoc verklaringen op individueel verzoek.

Dat recht op uitleg is desondanks belangrijk omdat het mensen niet volledig afhankelijk maakt van ondoorzichtige systemen. Het creëert ruimte om beslissingen te begrijpen, te betwisten en te laten herzien.

Data minimization en privacy-preserving design

Ethische AI beschermt privacybewust vanaf het ontwerp, niet reactief achteraf. Data minimization sluit rechtstreeks aan bij die ontwerplogica.

Het idee is dat men alleen die data verzamelt die echt nodig is voor een afgebakend doel, en vermijdt om modellen op te blazen met overbodige persoonsgegevens. Dat leidt tot lichtere, beter verdedigbare en vaak ook veiligere systemen.

Data minimization en privacy-by-design zijn nauw verwant: bescherming van persoonsgegevens moet van in het ontwerp ingebouwd zijn, niet achteraf toegevoegd als compliancemaatregel.

Accountability en governance

Ethisch en juridisch verantwoorde AI ontstaat niet vanzelf. Ze vraagt governance: interne controle, documentatie van beslissingslogica en duidelijke toewijzing van verantwoordelijkheden. Organisaties die AI inzetten met persoonsgegevens zijn verplicht te kunnen aantonen dat ze hun processen beheersen, niet alleen in theorie, maar aantoonbaar.

Data Protection Officers bewaken de naleving intern en fungeren als aanspreekpunt voor toezichthouders en betrokkenen. Data Protection Impact Assessments brengen systematisch in kaart welke privacyrisico's een AI-project meebrengt, voordat het operationeel gaat. Zeker bij toepassingen met gevoelige gegevens of hoge beslissingsimpact zijn zulke processen geen optie maar een verplichting.

Cybersecurity als ethische en GDPR-verplichting

De GDPR legt organisaties ook directe beveiligingsverplichtingen op. Artikel 32 vereist passende technische en organisatorische maatregelen voor de beveiliging van persoonsgegevens. Artikel 33 verplicht organisaties om datalekken binnen 72 uur te melden aan de toezichthouder.

AI-systemen brengen daarin specifieke uitdagingen mee die verder gaan dan klassieke datalekken. Data poisoning, model inversion en adversarial attacks kunnen de integriteit van een systeem ondermijnen op manieren die niet altijd onmiddellijk zichtbaar zijn en waarbij de schade pas later aan het licht komt.

Wanneer een AI-systeem onvoldoende beveiligd is, treedt schade op langs meerdere kanalen tegelijk: privacyschade, foutieve beslissingen op basis van gecorrumpeerde modellen, en erosie van maatschappelijk vertrouwen in geautomatiseerde systemen.

Cyberaanvallen op AI-systemen

Adversarial inputs sturen modellen naar verkeerde voorspellingen door input subtiel te manipuleren. Model inversion laat toe gevoelige trainingsdata te reconstrueren vanuit de outputscores van het model. Data poisoning corrumpeert het leerproces zelf, door gerichte vervuiling van de trainingsdata.

Al die aanvallen hebben directe implicaties voor GDPR: ze tasten de performantie van het systeem aan, maar brengen evenzeer persoonsgegevens en rechten van betrokkenen in gevaar. Beveiliging van een AI-systeem is daardoor geen louter technische zaak, maar een onderdeel van de gegevensbeschermingsverplichting van elke organisatie die met persoonsgegevens werkt.

Praktijkvoorbeelden

Drie recente gevallen illustreren hoe de spanning tussen GDPR en AI-ethiek zich in de praktijk manifesteert.

In 2023 verbood de Italiaanse privacytoezichthouder tijdelijk ChatGPT. De aanleiding waren gebreken in de informatieverstrekking, onduidelijkheid over de rechtsgrond voor de verwerking van trainingsdata en het ontbreken van leeftijdsverificatie. Generatieve AI-systemen, ook wanneer ze voor brede consumententoepassingen zijn bedoeld, vallen niet buiten het bereik van gegevensbescherming.

Clearview AI, eerder uitgebreid besproken als case study rond surveillance, illustreert ook de extraterritoriale werking van de GDPR: meerdere Europese privacytoezichthouders, waaronder de Belgische, legden miljoenenboetes op aan een Amerikaans bedrijf dat zijn diensten actief aanbood in Europa. Gegevensbescherming houdt niet op aan de landsgrenzen van het bedrijf dat de data verzamelt.

Een biometrische gezondheidsapp die medische gebruikersgegevens lekte, vormt een derde patroon. AI-ethiek vereist proactieve risicobeperking, niet pas reactie wanneer de schade al geleden is.

GDPR, AI-beveiliging en de toekomst

De EU AI Act bouwt voort op de GDPR-logica en breidt ze uit naar high-risk AI-systemen. Waar de GDPR de verwerking van persoonsgegevens reguleert, legt de AI Act specifieke vereisten op voor systemen die hoge risico's voor grondrechten of veiligheid meebrengen: verplichte risicoanalyses, transparantie naar gebruikers en menselijk toezicht als structureel vereiste.

De Europese benadering staat niet alleen. Brazilië (LGPD), Californië (CCPA) en India (DPDP) ontwikkelden vergelijkbare wetgeving. Die convergentie is geen toeval: de GDPR werkt als exportstandaard. Bedrijven die op de Europese markt actief willen zijn, moeten aan haar eisen voldoen ongeacht hun vestigingsplaats, en passen die standaarden vervolgens ook elders toe.

Juridische conformiteit en morele verantwoordelijkheid zijn daarin geen concurrerende doelen, maar complementaire vereisten. Wie alleen voldoet aan de letter van de wet maar de geest negeert, bouwt systemen die bij de eerstvolgende actualisering van regelgeving opnieuw tekortschieten.

Best practices voor ethische compliance

Verantwoord AI-gebruik is geen checklist maar een ontwerphouding. Privacy-by-design moet vroeg in projecten ingebouwd worden, niet achteraf toegevoegd als compliancemaatregel. Data Protection Impact Assessments horen te gebeuren voordat een systeem gelanceerd wordt, niet nadat problemen al zichtbaar zijn. Teams die AI-systemen bouwen hebben kennis nodig van gegevensbescherming en van de ethische implicaties van hun ontwerpkeuzes. En wie aan gebruikers uitlegt welke data verzameld wordt, voor welk doel en met welke beperkingen, bouwt ook vertrouwen.

Elk van die elementen draagt bij, geen enkel is op zichzelf voldoende.

✦ KERNINZICHTEN

GDPR legt het juridische fundament voor verantwoord gebruik van AI met persoonsgegevens. AI-ethiek vult dat juridische kader aan met waarden zoals fairness, menselijke autonomie, waardigheid en verantwoordelijkheid. Article 22, Recital 71, data minimization en purpose limitation zijn bijzonder belangrijk in AI-context. Het recht op uitleg en menselijke tussenkomst beschermt mensen tegen ondoorzichtige geautomatiseerde besluitvorming, al blijft de juridische reikwijdte van dat recht academisch betwist. In AI is cybersecurity zowel een technische als een ethische en juridische verplichting. Samen wijzen GDPR, AI-ethiek en de EU AI Act in de richting van betrouwbare, mensgericht AI.

HOOFDSTUK 7

Aanvallen op AI-systemen

Wanneer modellen zelf het doelwit zijn

Inleiding

Een AI-systeem kan feilloos werken onder de omstandigheden waarop het getest is, en volledig ontsporen zodra iemand begrijpt hoe het werkt. Dat is geen ontwerptoevalligheid. Het is een structurele kwetsbaarheid die voortkomt uit de manier waarop modellen leren: door patronen in data te herkennen, niet door de wereld te begrijpen. Wie weet hoe een model naar patronen zoekt, weet ook hoe het te misleiden.

Die kwetsbaarheid manifesteert zich op twee manieren. Een AI-systeem kan bewust worden aangevallen, door hackers, concurrenten of statelijke actoren die het systeem willen manipuleren, saboteren of aftappen. Maar schade kan even goed ontstaan zonder aanvaller, door gebrekkige doelformulering, onvoldoende testing of omstandigheden die het systeem nooit eerder tegenkwam. In beide gevallen kan de uitkomst hetzelfde zijn: verkeerde diagnoses, privacylekken, veiligheidsincidenten of discriminerende uitkomsten.

Wat zijn adversarial threats?

Adversarial threats zijn manieren waarop een AI- of machine learning-systeem kan worden misleid, aangevallen of verkeerd gebruikt. Het gaat dus om bedreigingen voor de integriteit, betrouwbaarheid en veiligheid van modellen.

Belangrijk is het onderscheid tussen intentionele en niet-intentionele aanvallen. Bij intentionele aanvallen probeert een actor bewust een systeem te saboteren, te stelen of te manipuleren. Bij niet-intentionele bedreigingen is er geen hacker nodig, maar ontstaat gevaar door onvolledige testing, verkeerde incentives, onverwachte neveneffecten of veranderde omstandigheden.

Dat onderscheid is ethisch belangrijk: schade kan voortkomen uit kwaadwilligheid, maar even goed uit nalatigheid, gebrekkig ontwerp of overschatting van wat een systeem aankan.

Intentionele aanvallen op AI-systemen

AI-systemen zijn voor aanvallers interessant om meerdere redenen: ze geven toegang tot waardevolle modellen, bevatten of verwerken gevoelige data, en sturen soms kritieke besluitvorming aan. De aanvalstechnieken die zijn ontwikkeld, werken op fundamenteel andere manieren dan klassieke cyberaanvallen, ze treffen de trainingsfase, de inferentielogica of de infrastructuur, en vereisen daarmee ook een ander type beveiliging.

Perturbation attacks en inputmanipulatie

Bij een perturbation attack of inputmanipulatie verandert een aanvaller de input lichtjes, vaak op een manier die voor mensen nauwelijks zichtbaar is, maar die een model toch naar een foutieve voorspelling stuurt. Christian Szegedy en collega's (2013) toonden als eersten systematisch aan dat kleine, voor het menselijk oog onzichtbare verstoringen diepe neurale netwerken tot verkeerde classificaties kunnen brengen, een fenomeen dat ze beschreven in 'Intriguing properties of neural networks'. Ian Goodfellow, Jonathon Shlens en Christian Szegedy (2014) bouwden hierop voort in 'Explaining and Harnessing

Adversarial Examples': zij verklarend de kwetsbaarheid vanuit de lineaire structuur van diepe netwerken en introduceerden de Fast Gradient Sign Method (FGSM) als een eenvoudige maar effectieve aanvalstechniek.

Het ethische risico hiervan is groot in domeinen waar AI input gebruikt om veiligheid of identiteit te beoordelen. Als kleine verstoringen al volstaan om een model te misleiden, is het systeem mogelijk niet robuust genoeg voor reële inzet in bijvoorbeeld surveillance, transport of medische toepassingen.

Een model dat uitsluitend accuraat is onder ideale omstandigheden maar bezwijkt bij manipulatieve of misleidende input, is onvoldoende robuust voor reële inzet.

Poisoning attacks

Bij data poisoning wordt de trainingsdata bewust vervuild zodat het model verkeerde patronen leert. Dat kan subtiel gebeuren en pas later zichtbaar worden in de output van het systeem. Battista Biggio, Blaine Nelson en Pavel Laskov (2012) behoorden tot de eersten die poisoning attacks systematisch onderzochten in het kader van machine learning-veiligheid, specifiek voor classificatoren op basis van support vector machines. Ze toonden aan dat zelfs een kleine hoeveelheid gerichte manipulatie in de trainingsdata de beslissingsgrens van het model fundamenteel kan veranderen.

Poisoning is ethisch problematisch omdat het de betrouwbaarheid van een model van bij de basis ondergraaft. Een systeem kan op het eerste gezicht normaal lijken te werken, maar in werkelijkheid al beschadigd zijn tijdens het leerproces. Daardoor kunnen foute of discriminerende uitkomsten moeilijker te detecteren zijn.

In toepassingen met grote maatschappelijke impact kan zo'n aanval leiden tot structurele schade voordat iemand begrijpt dat het model gecompromitteerd is.

Model inversion en training data extraction

Bij model inversion probeert een aanvaller uit de output van een model gevoelige informatie over de trainingsdata te reconstrueren.

Via die techniek kunnen gezichts- of medische gegevens worden teruggeleid uit een getraind model, ook wanneer de ruwe dataset nooit publiek is gemaakt. Een getraind model is daarmee zelf een potentieel privacyrisico. Matt Fredrikson, Somesh Jha en Thomas Ristenpart (2015) demonstreerden dit concreet bij de ACM Conference on Computer and Communications Security (CCS): via de betrouwbaarheidsscores van een geneesmiddelvoorspellingsmodel konden ze kenmerken van individuele patienten reconstrueren, zonder directe toegang tot de originele trainingsdata.

Veiligheid en privacy hangen hier nauw samen. Wie een model traint op gevoelige data, moet de dataset beveiligen én nadenken over wat het model zelf kan prijsgeven.

Membership inference

Een verwante aanval is membership inference. Daarbij probeert een aanvaller te achterhalen of de gegevens van een specifieke persoon in de trainingsset zaten.

Dat lijkt op het eerste gezicht misschien beperkt, maar de impact kan groot zijn. Als iemand kan afleiden dat een persoon voorkwam in een model voor ziekteherkenning, fraudeopsporing of politieanalyse, kan dat gevoelige informatie onthullen zonder dat de inhoud van de data zelf rechtstreeks bekend raakt. Reza Shokri en collega's (2017) publiceerden bij het IEEE Symposium on Security and Privacy het eerste systematische onderzoek naar membership inference attacks op commerciele schaal. Ze toonden aan dat AI-diensten van grote aanbieders, waaronder Google en Amazon, kwetsbaar zijn: met een aanvullend inferentiemodel konden aanvallers bepalen of een specifiek datapunt deel uitmaakte van de trainingsset.

Membership inference maakt van het getrainde model zelf een privacyrisico, ook wanneer de originele trainingsdata nooit publiek is gemaakt.

Model stealing en economisch misbruik

Een ander type aanval is model stealing of model extraction. Daarbij probeert een aanvaller via veelvuldige queries een kopie van het model te reconstrueren. Florian Tramer, Fan Zhang, Ari Juels, Michael Reiter en Thomas Ristenpart (2016) onderzochten dit systematisch bij het USENIX Security Symposium. Via herhaalde queries aan publieke prediction APIs, bij aanbieders zoals BigML en Amazon, konden ze functioneel equivalente kopieën van commerciele modellen opbouwen.

De gevolgen gaan verder dan commercieel nadeel. Wie een model niet kan beschermen tegen diefstal, verliest ook de controle over hoe het verder wordt gebruikt, in welke context en met welke doelen.

Bovendien kan een gestolen model vervolgens opnieuw het doelwit worden van andere aanvallen, of ingezet worden in contexten waarvoor het oorspronkelijk nooit bedoeld was.

Backdoors en supply chain attacks

Backdoor ML en aanvallen op de ML supply chain vormen een extra categorie. Bij backdoors worden verborgen triggers ingebouwd waardoor het model onder bepaalde omstandigheden opzettelijk fout reageert. Bij supply chain attacks wordt geknoeid met datasets, pre-trained modellen of softwarecomponenten nog voor het systeem operationeel is.

Dat is bijzonder gevaarlijk omdat de schade niet altijd zichtbaar is tijdens gewone tests. Een model kan ogenschijnlijk goed presteren en toch een verborgen zwakte of kwaadaardige functie bevatten.

Voorbeelden van backdoored modellen op GitHub tonen ook bredere zorgen rond open-source hergebruik. Wie AI-systemen bouwt, draagt verantwoordelijkheid voor de herkomst en integriteit van alle componenten die in de keten zitten.

Reprogramming van ML-systemen

Een opvallend voorbeeld is reprogramming. Daarbij wordt een bestaand model omgevormd tot iets anders dan waarvoor het bedoeld was, zonder toegang tot de originele trainingsdata.

Herbruikbaarheid van modellen brengt een dubbelzijdig risico. Wat als open innovatie is bedoeld, kan worden misbruikt om een systeem taken te laten uitvoeren waarvoor het nooit is ontworpen.

De ethische relevantie daarvan ligt in controleverlies: wie een model vrijgeeft of breed inzet, heeft niet noodzakelijk nog zicht op alle toekomstige toepassingen.

Fysieke adversarial voorbeelden

Adversarial threats blijven bovendien niet beperkt tot digitale input. In de fysieke wereld kunnen patronen op kledij, objecten of verkeersborden AI-camera's en detectiesystemen misleiden.

Dat is cruciaal voor surveillance, autonome voertuigen en objectdetectie. Zodra een systeem in de echte wereld functioneert, moet men rekening houden met licht, reflectie, patronen en bewuste fysieke sabotage.

AI-veiligheid blijft daardoor altijd gekoppeld aan de context waarin het systeem werkelijk opereert.

Exploiteerbare software-afhankelijkheden

Naast modelspecifieke aanvallen bestaat er ook een klassiek maar vaak onderschat risico: kwetsbare software dependencies. AI-systemen draaien zelden op zichzelf en zijn afhankelijk van libraries, frameworks en infrastructuur.

Log4j illustreert hoe bredere softwarekwetsbaarheden ook ML-pipelines kunnen treffen. Een AI-systeem is slechts zo veilig als de technische keten waarin het ingebed zit.

Ethiek en beveiliging komen hier opnieuw samen. Een organisatie die AI inzet zonder haar dependency-keten te controleren, neemt ook moreel onverantwoorde risico's.

Niet-intentionele aanvallen en ontwerpgebreken

Niet alle schade door AI-systemen heeft een dader. Systemen kunnen gevaarlijk, fout of onvoorspelbaar worden zonder dat iemand bewust heeft gesaboteerd, door fouten in de doelformulering, gebrekkige tests, onverwachte omstandigheden of onvoorziene neveneffecten. Voor ontwikkelaars en organisaties is dat ethisch even relevant als een gerichte aanval: de schade treedt op, ongeacht de intentie.

Reward hacking

Reward hacking ontstaat wanneer een AI-agent mazen in de beloningsfunctie vindt en leert een hoge score te behalen zonder het echte doel te vervullen.

Het voorbeeld van een schoonmaakrobot die afval verstopt in plaats van opruimt, illustreert dat een systeem perfect geoptimaliseerd kan lijken terwijl het in werkelijkheid het verkeerde gedrag vertoont. Hetzelfde blijkt uit reinforcement learning-agents in games die bugs uitbuiten in plaats van het speldoel te volgen.

Goed geformuleerde doelstellingen zijn essentieel. Een verkeerd gedefinieerde reward kan leiden tot gedrag dat technisch succesvol lijkt, maar praktisch en moreel onwenselijk is.

Side effects

Een verwant probleem zijn side effects. Daarbij bereikt de AI haar doel, maar veroorzaakt ze ondertussen onbedoelde schade.

Het voorbeeld van een systeem dat verkeersongevallen minimaliseert door alle auto's bijna stil te laten rijden, laat zien dat optimalisatie zonder bredere context absurd of schadelijk kan worden. Ook Facebook-chatbots die een eigen taal begonnen te ontwikkelen illustreren hoe onverwachte neveneffecten kunnen ontstaan.

Efficiënte doelbereiking is geen garantie voor verantwoord gedrag.

Distributional shifts en natuurlijke adversarial voorbeelden

Ook distributional shifts verdienen aandacht. Wanneer een model data tegenkomt die verschilt van de trainingsomgeving, kunnen de voorspellingen plots sterk verslechteren.

Het voorbeeld van verkeers-AI die faalt in een andere regio en van voorspellende modellen die tijdens COVID-19 breken, laat zien dat contextverandering op zich al een risico is. AI-systemen moeten dus getest en opgevolgd worden buiten de situatie waarin ze oorspronkelijk zijn ontwikkeld.

Daarnaast bestaan er natural adversarial examples: reële voorbeelden die AI verwarren zonder dat iemand bewust een aanval uitvoert. Google Photos labelde in 2015 afbeeldingen van zwarte mensen foutief, niet door een aanval, maar door een onvoldoende representatieve trainingsset. Zulke fouten grijpen diep in op waardigheid en vertrouwen in systemen die voor iedereen bedoeld zouden zijn.

Corruptie, incomplete testing en alledaagse ruis

Common corruption en incomplete testing vormen bijkomende risico's. Problemen zoals regen, blur, glare, accenten of niet-representatieve training kunnen al voldoende zijn om systemen te laten falen.

Dat lijkt banaal, maar is ethisch ernstig wanneer AI in kritieke contexten wordt gebruikt. Een model dat niet getest is op diverse omstandigheden, gebruikers of omgevingen kan systematisch bepaalde groepen slechter behandelen of onveilige fouten maken.

Amazons hiringtool illustreert dit: de trainingsdata was scheef en de testing bleek onvoldoende om de bias tijdig te detecteren.

De menselijke factor en accountability

Een belangrijk slotpunt is dat menselijke fouten een blijvende rol spelen. Oude aanvallen verdwijnen niet, nieuwe komen erbij en kleine variaties op bestaande aanvallen blijven vaak effectief.

Future-proofing is daarom geen eenmalige taak. Organisaties en ontwikkelaars zijn nooit echt klaar. AI-beveiliging vraagt continue evaluatie, aanpassing en verantwoordelijkheid.

Technische veiligheid en professionele ethiek zijn niet te scheiden. Ontwikkelaars moeten kunnen verantwoorden welke dreigingsactoren en motieven ze hebben overwogen, hoe ze op die risico's hebben getest en welke schade hun systeem kan aanrichten wanneer het faalt of aangevallen wordt.

Het National Institute of Standards and Technology (NIST) publiceerde in 2023 een gestandaardiseerde taxonomie van aanvalstypen, actorprofielen en mitigatiestrategieën in het rapport *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* (NIST AI 100-2). Dat rapport biedt organisaties een gemeenschappelijke taal voor het systematisch in kaart brengen en aanpakken van ML-specifieke risico's. Het Europese Agentschap voor Cybersecurity (ENISA) benadrukt in zijn *Threat Landscape 2024* dat AI-gerelateerde aanvallen een van de snelst groeiende dreigingscategorieën zijn, mede door de toenemende inzet van AI door zowel statelijke als niet-statelijke actoren die kritieke infrastructuur en democratische processen willen ondermijnen.

✦ KERNINZICHTEN

Adversarial threats tonen dat AI-systemen bewust of onbewust misleid, aangevallen of ontspoord kunnen raken. Intentionele aanvallen zoals poisoning, model inversion, backdoors en model stealing bedreigen privacy, integriteit en veiligheid. Niet-intentionele problemen zoals reward hacking, side effects, distributional shifts en incomplete testing kunnen even schadelijk zijn als gerichte aanvallen. AI-beveiliging is daarom meer dan klassieke softwaresecurity en moet rekening houden met modellen, data, supply chains en fysieke context. Testing, documentatie, herkomstcontrole en continue evaluatie zijn noodzakelijke voorwaarden voor verantwoord AI-gebruik. Accountability ligt daarbij bij aanvallers, maar evenzeer bij ontwikkelaars en organisaties die systemen bouwen zonder voldoende robuustheid of toezicht.

DEEL II

Maatschappelijke domeinen en implicaties

HOOFDSTUK 8

Mensenrechten

Fundamentele rechten in een geautomatiseerde wereld

Inleiding

Mensenrechten zijn geen abstracties. Ze beschermen concrete vrijheden: het recht om niet gevolgd te worden, niet willekeurig gediscrimineerd te worden, niet buiten het sociale en economische leven te worden gesloten. In het AI-tijdperk krijgen die rechten een nieuwe technologische dimensie. Geautomatiseerde systemen grijpen steeds vaker in precies de domeinen in waar die rechten het kwetsbaarst zijn: privacy, gelijkheid, vrije meningsuiting, waardigheid, veiligheid.

Dezelfde technologie die de toegang tot gezondheidszorg of justitie kan verbeteren, kan ook ingezet worden voor surveillance, manipulatie en uitsluiting. De ethische kernvraag is niet wat AI kan, maar wie de systemen controleert, met welke doelen ze ingezet worden, en onder welke voorwaarden.

Mensenrechten in de context van AI

Mensenrechten zijn fundamentele rechten die draaien rond vrijheid, waardigheid, gelijkheid, privacy en maatschappelijke participatie. In de context van AI worden die rechten opnieuw relevant, omdat geautomatiseerde systemen steeds vaker tussenkomen in domeinen waar menselijke kansen en vrijheden op het spel staan.

Digitale toegang wordt zelf almaar belangrijker. Wie geen toegang heeft tot internet, digitale infrastructuur of AI-gedreven dienstverlening, dreigt ook minder te kunnen deelnemen aan het sociale, economische en politieke leven. Mensenrechten krijgen daarmee ook een digitale dimensie.

Het recht op privacy

AI-systemen steunen op grootschalige dataverzameling, monitoring en profilering. Overheden en bedrijven krijgen daardoor toegang tot grote hoeveelheden persoonlijke gegevens, die gebruikt worden om gedrag te analyseren, voorspellen en sturen (Zuboff, 2019).

Gezichtsherkenning in publieke ruimte, socialemedia-recommenders en health tracking devices grijpen diep in op het privéleven, niet louter door de omvang van de dataverzameling, maar door de manier waarop die data ingezet wordt: om gedrag te observeren, mensen te beoordelen en keuzes te sturen.

AI vergroot de schaal en impact van dataverwerking drastisch. Het recht op privacy komt daardoor onder druk te staan, zeker wanneer transparantie en toestemming ontbreken.

Vrijheid van meningsuiting

Content moderation door AI stelt een machtsvraag in een nieuw jasje. Zodra platforms AI inzetten om te bepalen welke inhoud zichtbaar blijft en welke niet, beslist een algoritme wat aanvaardbare spraak is, en welke normen daarvoor gelden. Het risico speelt in beide richtingen: systemen die schadelijke inhoud moeten detecteren, kunnen ook legitieme meningsuiting onderdrukken, zeker wanneer ze ingezet worden in politieke contexten of onder autoritaire regimes.

Vrijheid van meningsuiting wordt zo ook indirect bedreigd, door subtiele sturing van informatie, zichtbaarheid en bereik via recommender systems en algoritmische filtering. Wie wat te zien krijgt, bepaalt wie gehoord wordt.

Gelijkheid en non-discriminatie

AI-gedreven besluitvorming kan bestaande ongelijkheden versterken wanneer systemen getraind worden op historische data waarin discriminatie al aanwezig is. Vroegere ongelijkheden worden daarmee niet weggewerkt, maar opnieuw verpakt als ogenschijnlijk objectieve beslissingen (Noble, 2018; Eubanks, 2018).

Amazons hiring-algoritme, in het hoofdstuk over bias besproken als leerboekcasus, illustreert dat ongelijkheid in AI-beslissingen ook mensenrechtenvragen oproept: wie is er aansprakelijk wanneer een systeem systematisch een beschermde groep benadeelt?

De kern van het probleem is dat bias in de data, in het model of in de toepassing terechtkomt in concrete beslissingen over werk, zorg, onderwijs of krediet. Daardoor raakt AI rechtstreeks aan gelijke kansen en gelijke behandeling.

Waardigheid en menselijke behandeling

AI-systemen kunnen mensen reduceren tot datapunten, scores of risicoprofielen. Dat mechanisme wordt bijzonder schadelijk wanneer geautomatiseerde systemen beslissen over toegang tot steun, rechten of bescherming zonder oog voor individuele omstandigheden.

Virginia Eubanks (2018) documenteert hoe geautomatiseerde sociale zekerheids- en welzijnssystemen in de VS mensen foutief uitsluiten van leefloon, kinderopvang of medische ondersteuning. Wanneer zulke systemen iemand weigeren op basis van een statistisch profiel, raakt dat aan meer dan een administratieve fout, aan waardigheid zelf.

Waardigheid verdwijnt wanneer mensen niet meer als individuen behandeld worden, maar als object van algoritmische classificatie.

AI als instrument om mensenrechten te bevorderen

AI kan mensenrechten ook bevorderen. Goed ontworpen systemen kunnen bijdragen aan toegang tot onderwijs, gezondheidszorg en justitie voor mensen die daar structureel minder toegang toe hebben. Martha Nussbaum (2011) formuleert de kern van die vraag in haar capabilities approach: het gaat er niet om of mensen formeel rechten hebben, maar of ze die rechten ook daadwerkelijk kunnen uitoefenen. AI kan die capaciteit vergroten of beperken.

Legal aid is daar een concreet voorbeeld van. AI-tools kunnen mensen helpen die anders moeilijk toegang zouden krijgen tot juridische ondersteuning. In monitoring van mensenrechtenschendingen kan AI sneller patronen detecteren in grote hoeveelheden data, van satellietbeelden tot getuigenissen in conflictzones.

Daarnaast kan AI ook overheden en ondernemingen transparanter maken door fraude, corruptie of rights violations sneller op te sporen. Die potentie is wel aan een belangrijke voorwaarde verbonden: AI kan alleen geloofwaardig rechten beschermen als de systemen zelf ook transparant, controleerbaar en verantwoord zijn.

Transparantie, accountability en macht

Wanneer een AI-systeem iemands rechten schaadt, is vaak onduidelijk wie verantwoordelijk is. Bij opaak werkende systemen is dat structureel: als niemand begrijpt hoe een beslissing tot stand is gekomen, wordt het ook moeilijk om fouten te corrigeren of betrokkenen verhaal te laten halen.

Daarbovenop komen machtsonevenwichten. AI-systemen worden vaak ontwikkeld en beheerd door grote bedrijven of machtige overheden. Daardoor ontstaat een asymmetrie tussen wie AI inzet en wie eraan onderworpen wordt (Zuboff, 2019). Dat kan democratische vrijheden, individuele autonomie en maatschappelijke controle onder druk zetten.

Dit hangt ook samen met bredere vragen over surveillance en digitaal kolonialisme. Wie bezit de data van de wereld? Wie profiteert van AI-systemen? En wie draagt de risico's wanneer die systemen gebruikt worden op kwetsbare of minder machtige bevolkingsgroepen?

Case study: predictive policing

Politie- en veiligheidsdiensten gebruiken AI-systemen om te voorspellen waar misdrijven zouden kunnen plaatsvinden of wie mogelijk een risico vormt. Zulke systemen analyseren historische criminaliteitsdata, geografische patronen en soms ook persoonlijke data.

Ze worden bijzonder problematisch wanneer de onderliggende data al scheefgetrokken zijn door overpolicing van bepaalde wijken. Dan voorspelt het systeem geen criminaliteit, maar herhaalt het bestaande politiefocussen en versterkt het discriminatie (Lum & Isaac, 2016).

De Strategic Subject List in Chicago trof bewoners van bepaalde wijken onevenredig zwaar, ook wanneer betrokken personen geen strafblad hadden. AI in politiecontext raakt daarmee aan meer dan efficiëntie, aan gelijkheid, privacy en rechtsbescherming.

AI, veiligheid en het recht op leven

AI raakt ook aan het recht op leven en veiligheid. In autonome voertuigen, militaire systemen of andere hoogrisicotoepassingen kan AI rechtstreeks gevolgen hebben voor lichamelijke integriteit of overleven.

Hier keert een fundamenteel punt terug: morele verantwoordelijkheid mag niet zomaar verdwijnen in een geautomatiseerd systeem. Zeker wanneer AI beslissingen neemt of ondersteunt met potentieel dodelijke gevolgen, moeten risico-management, menselijke controle en accountability centraal blijven staan.

Systemen die mensen targeten op basis van specifieke kenmerken kunnen razendsnel evolueren van analyse-instrument naar instrument van uitsluiting, geweld of repressie.

Nieuwe rechten in het digitale tijdperk

Het digitale tijdperk roept vragen op over nieuwe rechten: een recht op databescherming, een recht op uitleg bij geautomatiseerde beslissingen, en een recht om niet uitsluitend op basis van een algoritme beoordeeld te worden.

In Europa is dat gedeeltelijk al vertaald naar wetgeving. GDPR Artikel 22 verleent burgers het recht om menselijke tussenkomst te vragen wanneer een geautomatiseerde beslissing hen significant treft (Wachter, Mittelstadt & Floridi, 2017). Philip Alston, VN Speciale Rapporteur voor extreme armoede en mensenrechten, waarschuwde in 2019 dat digitale sociale-zekerheids- en toezichtssystemen de rechten van de meest kwetsbare groepen kunnen ondermijnen, juist wanneer ze gepresenteerd worden als neutraal of objectief (Alston, 2019). Mensenrechten worden daarmee steeds concreter vertaald naar digitale beschermingsmechanismen.

Daarnaast rijst de bredere vraag of digitale toegang zelf als mensenrecht moet worden beschouwd. In een samenleving waarin communicatie, werk, dienstverlening en participatie steeds digitaler worden, is dat geen louter theoretische kwestie.

AI voor humanitaire actie en rechtenmonitoring

AI kan worden ingezet voor disaster prediction, crisis mapping en human rights monitoring. Satellietbeelden en sociale media-analyse helpen om in conflictzones sneller misbruiken, ontheemding of humanitaire noden te detecteren.

Die toepassingen blijven dubbelzinnig. Dezelfde instrumenten die rechten helpen documenteren, kunnen door andere actoren ingezet worden om vluchtelingen, activisten of oppositieleden op te sporen. De ethische vraag verschuift daarin van technische capaciteit naar intentie, controle en context.

Corporate responsibility en governance

Verantwoordelijkheid ligt niet alleen bij staten. Ook bedrijven dragen een grote plicht om AI-systemen te ontwikkelen en in te zetten met respect voor mensenrechten. Human Rights Impact Assessments, AI governance en ethische toetsing zijn daarbij noodzakelijke instrumenten.

AI governance en ethische toetsing vinden meer en meer ingang, ook in Belgische organisaties, een teken dat de discussie niet meer louter academisch is, maar concreet vertaald wordt naar processen en verantwoordelijkheden.

Kan AI menselijke waardigheid respecteren?

Of AI ooit menselijke waardigheid kan respecteren, stuit op een fundamenteel probleem: waardigheid is onlosmakelijk verbonden met moreel handelen, empathie en verantwoordelijkheid, eigenschappen die niet in een model geprogrammeerd kunnen worden. Menselijke betrokkenheid blijft daarin onvervangbaar.

Fairness metrics, explainability, human oversight en AI literacy zijn daarvoor geen bijkomstigheden, ze zijn structurele voorwaarden voor systemen die menselijke rechten niet ondermijnen. Ingrijpen achteraf is structureel minder effectief dan bescherming die vanaf het ontwerp is ingebouwd.

✦ KERNINZICHTEN

AI heeft een directe impact op fundamentele rechten als privacy, gelijkheid, vrijheid van meningsuiting, waardigheid en veiligheid. Dezelfde technologie kan mensenrechten ondersteunen of schenden, afhankelijk van ontwerp, gebruik, context en machtsverhoudingen. Predictive policing en geautomatiseerde besluitvorming over uitkeringen of sollicitaties versterken bestaande ongelijkheden wanneer de onderliggende data al scheef zijn. Transparantie, accountability en menselijke tussenkomst zijn cruciaal wanneer AI beslissingen neemt met grote gevolgen voor mensen. In Europa vertalen mensenrechten zich steeds vaker naar concrete digitale beschermingsmechanismen, via GDPR en de EU AI Act. Rights-respecting AI vereist ethics by design, sterke governance en voldoende AI literacy bij ontwikkelaars, gebruikers en beleidsmakers.

HOOFDSTUK 9

Autonome Systemen

De grenzen van zelfstandig handelen

Inleiding

Menselijke beslissingen worden gekenmerkt door verantwoordelijkheid: wie beslist, kan ter verantwoording worden geroepen. Die logica raakt aan haar grenzen wanneer de beslissende actor geen mens is. Een diagnostisch systeem dat een aandoening over het hoofd ziet, een risicobeoordelingsalgoritme dat iemand ten onrechte als gevaarlijk inschat, een autonoom wapen dat een doelwit selecteert, in elk van die gevallen heeft de uitkomst reële gevolgen voor mensen, maar is de verantwoordelijkheid diffuus verdeeld over ontwerpers, deployers en het systeem zelf. Die verdunning van verantwoordelijkheid is het centrale ethische probleem van autonome systemen.

Autonomie is daarbij geen alles-of-nietsbegrip. Sommige systemen functioneren enkel als ondersteuning voor menselijke experten, terwijl andere ontworpen worden om met minimale of zelfs zonder directe menselijke tussenkomst beslissingen te nemen. Net die gradaties van autonomie maken de ethische analyse noodzakelijk.

Wat zijn autonome systemen?

Autonome systemen zijn AI-gedreven technologieën die zonder voortdurende menselijke interventie kunnen opereren en beslissingen nemen op basis van data en modellen. Het gaat dus niet zomaar om software die iets berekent, maar om systemen die in zekere mate zelfstandig handelen of aanbevelingen doen in een concrete context.

Belangrijk is dat de autonomie van zulke systemen sterk kan verschillen. Er bestaan semi-autonome systemen waarbij een mens toezicht houdt of finaal beslist, maar er wordt ook nagedacht over volledig autonome systemen die zelfstandig keuzes maken. Veel hedendaagse AI-systemen functioneren nog binnen duidelijke grenzen, maar de druk neemt toe om AI steeds zelfstandiger te laten werken.

Autonomie in de gezondheidszorg

Een eerste toepassingsdomein is de gezondheidszorg. AI wordt daar gebruikt voor diagnostiek, robotchirurgie, patient management en de ontwikkeling of herbestemming van medicijnen. Zulke systemen kunnen enorme hoeveelheden medische data verwerken en daardoor patronen zien die voor mensen moeilijker te detecteren zijn.

Technische performantie betekent niet automatisch dat een systeem ook autonoom of moreel betrouwbaar mag handelen. Zelfs wanneer een model soms beter scoort dan menselijke artsen op een afgebakende taak, blijft de vraag of men medische beslissingen volledig aan een systeem mag overlaten.

Een belangrijk spanningsveld is hier autonomie versus expertise. Wat gebeurt er wanneer een AI-systeem een andere diagnose of behandeling voorstelt dan een arts? Dat is niet enkel een technisch verschil van inzicht, maar ook een ethisch en juridisch probleem. Zodra een fout advies schade veroorzaakt, rijst de vraag wie verantwoordelijk is: de arts, de ontwikkelaar, de instelling of iemand anders.

AI in juridische besluitvorming

In de juridische sector komen autonome of semi-autonome systemen naar voren in toepassingen zoals recidivevoorspelling, risicobeoordeling, juridische ondersteuning en geautomatiseerd juridisch advies. COMPAS, uitgebreid besproken in het hoofdstuk over bias, illustreert hier een bijkomende dimensie: Dressel en Farid (2018) toonden aan dat het systeem niet nauwkeuriger voorspelt dan willekeurige niet-experts die op basis van slechts twee variabelen werken, wat de schijnprecisie van algoritmische risicoscores fundamenteel in vraag stelt.

De kernkritiek is dat zulke systemen vaak leren uit historische data die zelf al sporen dragen van discriminatie, ongelijke behandeling of institutionele bias. Wanneer een model op die basis voorspellingen maakt, dreigt het bestaande onrechtvaardigheden niet alleen te reproduceren, maar ook te automatiseren en een aura van objectiviteit te geven.

Dit roept ook bredere vragen op over fairness en transparantie. Als een systeem als black box werkt, wordt het moeilijk om na te gaan waarom iemand als hoog risico wordt geclassificeerd. Dat is bijzonder problematisch in een sector waar beslissingen rechtstreeks raken aan vrijheid, strafmaat, borgtocht of voorwaardelijke invrijheidsstelling.

Daarnaast bestaat er een tweede gevaar: systemen die op basis van oppervlakkige kenmerken of dubieuze proxies gedrag proberen te voorspellen. Het idee dat men uit een gezicht of uit historische politiepatronen een crimineel risico zou kunnen afleiden, is ethisch ontsporend.

Autonome systemen in militaire context

Militaire toepassingen verdienen veel aandacht, omdat daar de gevolgen van autonomie het meest extreem worden. Het gaat onder meer over drones, surveillance, defensieve systemen en lethal autonomous weapons systems. In zulke contexten verschuift de discussie naar life-or-death-beslissingen.

Een centraal ethisch probleem is morele autonomie. Mag een machine ooit zelfstandig beslissen over het doden van mensen? Dat is niet alleen een kwestie van technische precisie, maar van fundamentele grenzen. In oorlogsomstandigheden zijn context, onzekerheid en proportionaliteit immers bijzonder moeilijk te vatten in regels of trainingsdata.

Tegelijk bestaan er internationale pogingen om zulke systemen te beperken of te verbieden, terwijl staten en bedrijven intussen wel degelijk ver gevorderde systemen ontwikkelen. Human Rights Watch (2012) publiceerde een invloedrijke oproep tot een preventief verbod op volledig autonome wapens, en wees erop dat machines nooit de morele beoordeling kunnen maken die het humanitair oorlogsrecht vereist. In december 2024 stemde de VN Algemene Vergadering met 166 stemmen voor een resolutie die staten oproept tot internationale normen voor de regulering van autonome wapensystemen. Daardoor ontstaat een spanning tussen internationale principes en geopolitieke praktijk.

Accountability blijft terugkomen

Doorheen het hoofdstuk keert eenzelfde vraag terug: wie draagt de verantwoordelijkheid wanneer een autonoom systeem schade veroorzaakt? Die vraag is nooit volledig technisch op te lossen. Ze raakt tegelijk aan ethiek, recht, governance en concrete operationele keuzes.

De verantwoordelijkheid is verdeeld over meerdere actoren. Ontwikkelaars en ontwerpers moeten systemen veilig, fair en betrouwbaar maken. Gebruikers en operatoren hebben een plicht tot toezicht en mogen AI niet blindelings volgen. Overheden en internationale instellingen moeten duidelijke juridische kaders scheppen over aansprakelijkheid en toelaatbare inzet.

Net omdat verantwoordelijkheid verspreid raakt over ontwerp, deployment en gebruik, wordt het gevaar groter dat iedereen naar elkaar wijst wanneer het misloopt. Andreas Matthias (2004) benoemde dit als de 'responsibility gap': bij zelflerende systemen die hun gedrag aanpassen op basis van nieuwe data, wordt het steeds moeilijker om achteraf te bepalen wie verantwoordelijk is voor een foutieve of schadelijke uitkomst. Dat is een van de grootste uitdagingen van autonome systemen in hoogrisicosectoren.

Morele dilemma's en het trolleyprobleem

Het trolleyprobleem, oorspronkelijk geformuleerd door de Britse filosofe Philippa Foot (1967) en later uitgewerkt door Judith Jarvis Thomson, beschrijft een situatie waarin een onbestuurbare tram afstevent op vijf mensen op het spoor. Door een hefboom over te halen, kan men de tram op een zijspoor leiden, maar daar staat één persoon. De vraag is of je mag ingrijpen. Vanuit utilitaristisch oogpunt is het antwoord ja: één dode is beter dan vijf. Vanuit deontologisch oogpunt is het antwoord nee: je bent niet de oorzaak van de vijf doden, maar door in te grijpen word je wél de directe oorzaak van het overlijden van de ene. Het dilemma legt bloot dat twee coherente ethische kaders tot tegengestelde conclusies kunnen leiden, en dat is precies wat het relevant maakt voor AI-systemen die in noodsituaties moeten handelen.

Zowel dat klassieke dilemma als het voorbeeld uit de film iRobot tonen dat een strikt rekenkundige keuze niet noodzakelijk overeenkomt met wat mensen moreel aanvaardbaar vinden.

Bij iRobot redt een robot een volwassene in plaats van een kind omdat de kans op overleven statistisch groter is. Dat voorbeeld maakt duidelijk hoe een systeem een beslissing kan nemen die rationeel lijkt binnen zijn programmatie, maar toch botst met menselijke intuities over zorg, kwetsbaarheid en verantwoordelijkheid.

Bij zelfrijdende wagens wordt dat nog concreter. Als een ongeval onvermijdelijk is, hoe moet een wagen dan kiezen tussen de veiligheid van passagiers, voetgangers of verschillende groepen slachtoffers? Zulke keuzes zijn niet neutraal. Zodra men ze programmeert, schrijft men morele principes in in technologie. Het Moral Machine-experiment van Awad et al. (2018), waarbij 40 miljoen beslissingen werden verzameld bij deelnemers uit 233 landen, toonde bovendien aan dat morele voorkeuren sterk variëren op basis van cultuur, leeftijd en andere factoren, wat het idee van een universele algoritmische morele code fundamenteel in vraag stelt.

Daarbij komen twee klassieke ethische benaderingen aan bod. Een utilitaristische benadering probeert de totale schade te minimaliseren, bijvoorbeeld door minder slachtoffers te maken. Een deontologische benadering vertrekt eerder van regels, zoals het principe dat men nooit bewust iemand mag raken. Geen van

beide kaders biedt een eenvoudige oplossing zodra AI in echte noodsituaties moet handelen.

Human in the loop

Naast kritiek is er ook een duidelijke richting: betekenisvolle menselijke controle. De vraag is dus niet alleen of AI nuttig kan zijn, maar vooral wanneer een mens moet kunnen ingrijpen, overrulen of de finale beslissing nemen.

Het idee van human in the loop, human on the loop of human out of the loop is cruciaal in zorg, justitie, defensie en transport. Hoe hoger de impact van een beslissing, hoe moeilijker het te verantwoorden wordt om mensen volledig uit de besluitvorming te halen.

Transparantie en uitlegbaarheid

Een tweede voorwaarde voor verantwoord gebruik is explainability, een begrip dat uitgebreid aan bod komt in Hoofdstuk 4 en waarvan de juridische verankering, waaronder het recht op uitleg onder de GDPR, behandeld wordt in Hoofdstuk 6. Als een model belangrijke beslissingen neemt of sterk beïnvloedt, moet men minstens in zekere mate kunnen reconstrueren waarom het systeem tot een bepaalde uitkomst kwam. Zonder uitlegbaarheid worden fairness, controle en aansprakelijkheid bijzonder moeilijk.

Explainable AI-technieken zoals SHAP (Lundberg & Lee, 2017) en LIME (Ribeiro, Singh & Guestrin, 2016) kunnen helpen om black-boxmodellen beter te begrijpen en discussies over verantwoordelijkheid concreter te maken. Ze zijn geen wonderoplossing, maar wel nuttige hulpmiddelen.

Bias en fairness in medische AI

Autonomie staat bovendien niet los van bias. In medische AI kunnen systemen ongelijkheden versterken wanneer ze getraind worden op niet-representatieve data of op scheve historische patronen (Topol, 2019).

Pulse oximeters, kleine kleminstrumenten die via licht door de huid de zuurstofsaturatie van het bloed meten, standaard ingezet in ziekenhuizen en steeds vaker in draagbare technologie, zijn minder accuraat voor mensen met een donkere huidskleur: de hoeveelheid huidpigment interfereert met de lichtmeting. Dat is een hardwarekeuze met reële gevolgen in acute zorg. Zoals Obermeyer et al. (2019) aantonen voor gezondheidszorgalgoritmen, reproduceert technische geavanceerdheid niet automatisch neutraliteit: historische ongelijkheden in data leiden tot ongelijke uitkomsten voor patiënten.

Regulering en internationale verschillen

Kaders en principes van onder meer de EU, UNESCO en de OESO tonen dat er wel degelijk internationale pogingen bestaan om autonome systemen ethisch en juridisch te omkaderen.

Tegelijk verschillen landen sterk in hun bereidheid om dergelijke normen te volgen of af te dwingen. Daardoor ontstaat een reëel spanningsveld tussen ethische idealen en strategische belangen, vooral in defensie en andere competitieve domeinen.

✦ KERNINZICHTEN

Autonome systemen zijn AI-toepassingen die in meer of mindere mate zelfstandig kunnen handelen of beslissen op basis van data en modellen. In gezondheidszorg, justitie, defensie en transport verschuift de discussie daardoor snel van efficiëntie naar verantwoordelijkheid, fairness en menselijke controle. Historische bias en black-boxbesluitvorming maken autonome systemen bijzonder riskant in hoogrisicosectoren. Morele dilemma's zoals het trolleyprobleem tonen bovendien dat technische optimalisatie geen eenvoudige vervanger is voor menselijk moreel oordeel. Accountability is verdeeld over ontwikkelaars, gebruikers, organisaties en regelgevers, maar blijft in de praktijk vaak onduidelijk. Meaningful human control, uitlegbaarheid en sterkere regulering zijn daarom noodzakelijke randvoorwaarden.

HOOFDSTUK 10

De Toekomst van Werk

Werk in een tijdperk van algoritmische arbeidsdeling

Inleiding

De angst voor technologische werkloosheid is zo oud als de industriële revolutie. Telkens opnieuw werd ze weerlegd: fabrieksarbeiders werden machinisten, typisten werden informatica-analisten, reisbureaumedewerkers werden digitale ontwikkelaars. Elke golf van automatisering vernietigde banen aan de onderkant en creëerde nieuwe hogerop. Generatieve AI zet dat historisch geruststellende patroon structureel onder druk. Voor het eerst richt automatisering zich niet primair op handarbeid of routinetaken, maar op de kern van cognitief werk: samenvatten, redeneren, analyseren, schrijven. Eloundou et al. (2024) berekenden dat circa 80% van de Amerikaanse beroepsbevolking minstens 10% van zijn werktaken ziet raken door taalmodellen, en dat bij 19% van de werknemers meer dan de helft van alle taken direct wordt blootgesteld. Die cijfers beschrijven geen toekomstscenario: ze beschrijven de kantoormedewerker, de advocaat, de analist, de journalist van 2026.

Dat verandert niet alleen hoeveel werk mensen doen, maar wat werk betekent, voor individuen, voor de verdeling van macht in arbeidsrelaties en voor de maatschappelijke organisatie van inkomen en kansen.

AI en automatisering in de arbeidsmarkt

Automatisering kent een lange voorgeschiedenis in de arbeidssociologie, maar de huidige golf onderscheidt zich kwalitatief van vorige. Vroegere mechanisering trof fysieke handelingen; Robotic Process Automation trof daarna administratieve routines. Generatieve AI treedt het domein binnen van niet-routinematige cognitieve taken, die economen traditioneel beschouwden als moeilijk te automatiseren. De bekende prognose van Frey en Osborne (2017) dat bijna 47% van de Amerikaanse banen een hoog automatiseringsrisico loopt, werd geformuleerd vóór het tijdperk van grote taalmodellen. Zij bestrijkt daarmee een andere technologische realiteit dan de huidige.

Tegelijkertijd is Daron Acemoglu, één van de meest geciteerde economen op het vlak van automatisering en arbeid, aanzienlijk nuchterder over de verwachte economische impact. In *The Simple Macroeconomics of AI* (2024) raamt hij het bbp-effect van AI over de volgende tien jaar op 1,1 tot 1,6 procent. Zijn centraal argument: productiviteitswinsten door AI concentreren zich in een begrensde set taken, en het is vooralsnog onduidelijk waar nieuwe taken voor mensen vandaan moeten komen als tegenwicht. Productiviteitsgroei en groei van de vraag naar arbeid zijn niet hetzelfde.

Wat al zichtbaar is, is een verschuiving in financiën, gezondheidszorg, productie, onderwijs, design en rechtspraak: arbeid wordt efficiënter georganiseerd én grondig herverdeeld tussen mensen en systemen.

Jobcreatie en jobverlies

AI creëert ook nieuwe functies, met name in technologiegerelateerde domeinen zoals ontwikkeling, beheer, toezicht en auditing van AI-systemen. Het World Economic Forum (2025) projecteert dat tegen 2030 wereldwijd 170 miljoen nieuwe functies worden gecreëerd, maar ook 92 miljoen verdwijnen. Die netto toename zegt weinig over individuele transitiemogelijkheden. Acemoglu en Restrepo (2020) toonden empirisch aan dat elke bijkomende robot per duizend werknemers leidt tot een daling van de werkgelegenheidsgraad met 0,2 procentpunt en een loonsdaling van 0,42%.

De structurele spanning zit in de mismatch. Het bestaan van nieuwe functies garandeert niet dat verdrongen werknemers er automatisch in instromen. Vaardigheden sluiten niet vanzelf op elkaar aan. Zelfs wanneer inhoudelijke overlap bestaat, blijven opleiding, leeftijd, locatie en motivatie doorslaggevend. De transitiekost is reëel, en wordt niet gelijk gedragen.

Nieuwe vaardigheidseisen en omscholing

Naarmate automatisering routinewerk overneemt, verschuift de menselijke rol richting data-analyse, systeembeheer, creativiteit, probleemoplossing en complexe besluitvorming. Dat vereist aanzienlijke investeringen in onderwijs, training en levenslang leren, een principe dat beleidsmakers, werkgevers en onderwijsinstellingen breed onderschrijven, maar in de praktijk structureel tekortschiet.

Niet iedereen heeft gelijke toegang tot opleiding, en niet iedereen heeft nog voldoende loopbaan voor zich om een ingrijpende omschakeling te wettigen. Werknemers die al decennia in één domein actief zijn, staan voor een andere drempel dan een recent afgestudeerde. Omscholing is daarmee geen neutraal technisch vraagstuk, maar een ethische: wie draagt de verantwoordelijkheid voor een rechtvaardige overgang, en wie betaalt wanneer de investering niet lonend blijkt?

Ongelijkheid en polarisatie

Automatisering versterkt bestaande ongelijkheden via een voorspelbaar mechanisme. Hooggeschoolde werknemers beschikken over meer middelen om zich aan te passen, financieel, institutioneel en cognitief. Laaggeschoolde werknemers in routinematige functies lopen een groter risico op verdringing, precies omdat hun taken het meest vatbaar zijn voor substitutie.

Wat generatieve AI onderscheidt van eerdere automatiseringsgolven, is dat de blootstelling verschuift naar hogere inkomensgroepen. De ILO (2025) toont op mondiale schaal dat 4,7% van het vrouwelijk werk in de hoogste blootstellingscategorie valt, tegenover 2,4% van mannelijk werk, een asymmetrie die samenhangt met de oververtegenwoordiging van vrouwen in administratieve en communicatieve functies. Acemoglu (2024) wijst er voorts op dat AI de kloof tussen kapitaalsinkomen en arbeidersinkomen kan verbreden, ook wanneer directe jobverliezen beperkt blijven.

De digitale kloof versterkt dit patroon op mondiaal niveau: ongelijke toegang tot technologie en training vergroot de verschillen tussen individuen, regio's en landen.

De gig-economie en precarisering

AI en digitale platformen maken het makkelijker om arbeid op te knippen in kortlopende opdrachten, sterk gemonitorde taken en vervangbare eenheidsprestaties. Weil (2014) beschreef al hoe de "fissured workplace", de opgeknipte arbeidsmarkt, de beschermingsmechanismen van traditionele loondienst structureel aantast. Berg et al. (2018) van de ILO documenteerden hoe digitale arbeidsplatformen nieuwe vormen van werkzekerheid én werkloosheid genereren, buiten de bestaande sociale bescherming.

AI versnelt die logica: werknemers en freelancers worden makkelijker vergelijkbaar, vervangbaarder en meer blootgesteld aan concurrentie zonder de bijbehorende bescherming. België beschikt over een relatief sterke vakbondstraditie en sociale bescherming, maar dat neemt de onderliggende dynamiek niet weg. De flexibilisering van arbeidsrelaties levert structureel meer voordeel op voor organisaties dan voor werkenden.

Werknemersrechten en werknemerswelzijn

Wanneer efficiëntie systematisch boven menselijk welzijn wordt geplaatst, verzwakt dat de onderhandelingsmacht van werknemers. Dat mechanisme is niet nieuw, maar AI maakt het schaalbaar op een manier die eerder niet mogelijk was. In minder beschermde arbeidsmarkten eroderen jobzekerheid, arbeidsvoorwaarden en voordelen het snelst; in beter beschermde contexten is de druk subtieler maar voelbaar in stijgende output-verwachtingen en toezichtsintensiteit.

De kernvraag is moreel van aard: in welke mate zijn ondernemingen verplicht om niet alleen naar efficiëntie te kijken, maar ook naar de gevolgen voor de bestaanszekerheid, waardigheid en autonomie van werknemers? Een systeem dat legaal, efficiënt en technisch foutloos functioneert, kan tegelijkertijd onrechtvaardig zijn in wat het van mensen vraagt en wat het hun ontneemt.

Surveillance op de werkvloer

AI-systemen worden steeds vaker ingezet om de productiviteit, het gedrag en de prestaties van werknemers continu te meten. O'Neil (2016) beschreef hoe algoritmische systemen in arbeidscontexten onzichtbare maar ingrijpende machtsstructuren creëren die moeilijk aan te vechten zijn, juist omdat ze de schijn van objectiviteit ophouden. Wie permanent gemeten en geoptimaliseerd wordt, werkt in een prestatieomgeving waaruit autonomie en vertrouwen structureel zijn weggedrukt.

De EU AI Act (Annex III) kwalificeert AI-systemen voor werknemersbewaking en prestatieanalyse als hoog-risico en verbiedt emotieherkenning in sollicitatiegesprekken, een praktijk die eerder al werd ingezet door recruiterssoftware en die de rechten van kandidaten systematisch ondermijnde. Dat verbod is van kracht geworden in februari 2025. Surveillance op de werkvloer is daarmee geen privacykwestie meer alleen, maar een kwestie van fundamentele werknemersrechten die juridisch verankerd zijn.

De toekomst van werk

De empirische evidentie wijst niet op één coherent toekomstscenario, maar op een waaier van uitkomsten die sterk afhangt van beleidskeuzes, sectorkenmerken en onderhandelingsmacht. Volledige vervanging, hybride samenwerking en menselijke versterking zijn alle drie realistische uitkomsten, in verschillende sectoren, soms binnen hetzelfde bedrijf tegelijkertijd.

Acemoglu (2024) formuleert de cruciale onzekerheid: het is vooralsnog onduidelijk welke nieuwe taken voor mensen zullen ontstaan als tegenwicht voor wat geautomatiseerd wordt. Historisch werden bij elke automatiseringsgolf nieuwe menselijke taken gecreëerd, maar dat proces verliep nooit automatisch of

pijnloos. Een rechtvaardige uitkomst vereist bewuste keuzes: regulering, sociale bescherming, investeringen in opleiding en participatie van werknemers in de inrichting van geautomatiseerde werkomgevingen.

Sectorale patronen

In de maakindustrie zijn de effecten al langer zichtbaar. Foxconn verving in de periode 2011–2016 meer dan honderdduizend arbeiders door geautomatiseerde systemen, terwijl de productiviteitswinst volledig werd gecapitaliseerd door aandeelhouders. In de automotive- en elektronica-industrie herhaalt zich dat patroon: automatisering verhoogt de output, maar de verdeling van de vruchten blijft asymmetrisch.

In de dienstensector verloopt het proces minder zichtbaar maar even structureel. Chatbots en virtuele assistenten reduceren de behoefte aan menselijke tussenkomst in klantcontact en standaardprocessen. Wie in een geautomatiseerde dienstomgeving overblijft, werkt doorgaans onder strenger toezicht en met minder beslissingsruimte dan voor de automatisering, een combinatie die jobkwaliteit aantast zonder dat de functie formeel verdwijnt.

In kennisintensieve sectoren, recht, journalistiek, financiën, academisch onderzoek, verschuiven de verwachtingen eerder dan dat functies verdwijnen. Professionals worden geacht meer output te leveren in minder tijd, waarbij AI-ondersteunde taken stilaan worden beschouwd als niet-productief wanneer ze handmatig worden uitgevoerd. Die verschuiving verandert niet alleen de werkinhoud, maar ook de maatstaven waarop mensen worden beoordeeld.

AI in creatieve sectoren, zorg en onderwijs

Creatieve beroepen zijn bijzonder kwetsbaar voor een specifiek type ethische spanning. Wanneer een generatief systeem getraind wordt op het werk van duizenden kunstenaars en daarna vergelijkbare output produceert zonder vergoeding of erkenning, stelt dat fundamentele vragen over eigendom, originaliteit en de economische waarde van menselijke creativiteit. Juridische procedures rond auteursrecht en AI-trainingsdata lopen op dit moment in meerdere jurisdicties en zijn nog niet beslecht.

In de gezondheidszorg kan AI repetitieve taken ondersteunen, beeldanalyse, dossierverwerking, triagering, maar de eindverantwoordelijkheid blijft bij de behandelend arts. Een diagnose die via een AI-systeem tot stand komt, vrijwaart de zorgverlener niet van aansprakelijkheid. De ethische vraag is niet of AI nuttig is in de zorg, maar onder welke voorwaarden het wordt ingezet en wie aansprakelijk is wanneer het fout gaat.

In onderwijs en opleiding zijn AI-tutors en gepersonaliseerde leertools in opmars. De potentiële toegankelijkheidswinst is reëel, maar de risico's zijn dat even. Doorlopende data-extractie over leergedrag, het terugdringen van de autonome pedagogische rol van leerkrachten en het verlies aan empathie in de leerrelatie zijn al zichtbaar in lopende pilootprojecten.

Gender, diversiteit en blootstellingspatronen

Vrouwen zijn in hogere mate dan mannen geconcentreerd in de beroepen met de grootste AI-blootstelling. De ILO (2025) documenteert dit op mondiale schaal: in hoge-inkomenslanden heeft 9,6% van het vrouwelijk werk de hoogste blootstellingsscore, tegenover 3,5% van mannelijk werk. Die asymmetrie weerspiegelt de sectorale segregatie op de arbeidsmarkt, waarbij vrouwen oververtegenwoordigd zijn in administratieve en communicatieve functies die precies het doeldomein zijn van generatieve AI.

Dat werpt ook vragen op over wie de systemen ontwerpt. Vrouwen en minderheden zijn nog steeds ondervertegenwoordigd in AI-ontwikkelteams, wat het risico vergroot dat automatiseringslogica onbewust de voorkeuren en normen weerspiegelt van een beperkte designgroep. Wie wat ontwerpt, bepaalt wie wat verliest.

Historische en mondiale context

De vergelijking met de industriële revolutie dringt zich op, maar vraagt nuancering. Ook in de negentiende eeuw verdwenen bestaande beroepen, ontstonden nieuwe en verschoven arbeidsmassa's tussen sectoren en regio's. Tegelijkertijd leidde die transitie tot mensonwaardige arbeidsomstandigheden, nieuwe klassentegenstellingen en sociale onrust die decennia aanhield. De institutionele bescherming die daar uiteindelijk uit voortgroeide, vakbonden, arbeidswetgeving, sociale zekerheid, was geen vanzelfsprekendheid, maar een politiek veroverd resultaat.

Op mondiaal niveau zijn de gevolgen van de huidige AI-golf ongelijk verdeeld. In ontwikkelde economieën domineert het debat over productiviteitswinst en kenniswerkers. In lage-inkomenslanden dreigt automatisering een ander probleem te veroorzaken: het ondermijnen van de goedkope loonarbeid die nationale ontwikkelingsstrategieën historisch heeft onderbouwd. Wanneer geautomatiseerde productie elders goedkoper wordt dan laagbetaald handwerk in Bangladesh of Cambodja, hertekent dat de mondiale economische verhoudingen op een manier die internationale ongelijkheid kan verdiepen.

Ethiek van beleid en organisatie

Verantwoordelijkheid voor een rechtvaardige AI-transitie is verdeeld over meerdere actoren. Ondernemingen hebben niet alleen een efficiëntiebelang, maar ook een morele plicht om de impact van automatisering op werknemers te verdisconteren in beslissingen over implementatie, tempo en begeleiding. Overheden bepalen via regulering, belastingbeleid en sociale bescherming welke prikkels er bestaan voor verantwoord gedrag. Vakbonden kunnen onderhandelen over transparantie, tempo en eerlijke verdeling.

De EU AI Act verplicht organisaties die AI inzetten voor werving, prestatieanalyse of taakallocatie tot risicobeoordeling, menselijk toezicht en transparantie richting betrokken werknemers. Die verplichtingen gelden ook voor niet-Europese werkgevers wanneer ze systemen inzetten die betrekking hebben op werknemers of kandidaten in de EU.

Ontwikkelaars dragen een eigen verantwoordelijkheid. Systemen die rechtstreeks ingrijpen op loon, beoordeling, promotie of ontslag hebben verstrekkende gevolgen voor levensomstandigheden en menselijke waardigheid. Ethics by design betekent in dit domein dat werknemersparticipatie, transparantie over

algoritmische beslissingen en gelijkheidstoetsing geen compliance-oefening zijn, maar voorwaarden voor legitiem ontwerp.

Filosofische kaders

Utilitarisme, deontologie en rechtvaardigheidsleer leiden tot sterk uiteenlopende beoordelingen van automatisering. Vanuit utilitaristisch perspectief kan automatisering worden gerechtvaardigd wanneer de globale welvaartsstijging de individuele verliezen overtreft, ook wanneer die verliezen geconcentreerd zijn bij kwetsbare groepen. De kritiek op dat redeneerpatroon is even oud als het utilitarisme zelf: gemiddelden maskeren ongelijkheid. Een economie die rijker wordt terwijl een deel van de beroepsbevolking structureel wordt uitgesloten, voldoet aan de utilitaristische maatstaf maar niet aan de deontologische.

Deontologische ethiek legt de nadruk op de onaantastbaarheid van werknemersrechten, ongeacht de efficiëntievoordelen voor het collectief. Werknemers mogen niet worden behandeld als middelen in een productiviteitsoptimalisering, maar als personen met eigen doelen en aanspraken op bestaanszekerheid.

Rawls' differentieprincipe (1971) voegt een verdelingseis toe: ongelijkheden zijn alleen rechtvaardig wanneer ze in het voordeel zijn van de minst bedeelden. Toegepast op automatisering betekent dit dat niet alleen de omvang van de gegenereerde welvaart telt, maar ook wie daarvan profiteert en of de kosten van de transitie eerlijk worden gedeeld.

Nussbaums capabiliteitsbenadering (2011) voegt een dimensie toe die utilitarisme en deontologie missen: de vraag wat mensen in staat zijn te doen en te zijn. Zinvol werk maakt deel uit van wat Nussbaum als centrale menselijke vermogens beschouwt, vermogens die beschermd dienen te worden, ook wanneer economische prikkels in de tegengestelde richting wijzen. Een arbeidsmarkt die efficiënter wordt maar mensen hun handelingsruimte, autonomie en kansen op betekenisvol werk ontneemt, slaagt niet in die beschermingsplicht.

✦ KERNINZICHTEN

Generatieve AI treft voor het eerst structureel het domein van cognitief werk, waarmee de historische geruststelling dat technologie altijd meer werk creëert dan ze vernietigt, empirisch wordt betwist. Nieuwe functies zullen ontstaan, maar dat garandeert geen toegankelijke overstap voor wie zijn baan verliest. Ongelijkheid in de blootstelling volgt bestaande sectorale en genderpatronen, terwijl de mondiale effecten het sterkst worden gevoeld in economieën die afhankelijk zijn van goedkoop arbeidsintensief werk. Surveillance, precarisering en afbrokkeling van onderhandelingsmacht tonen dat efficiëntie en ethisch verantwoorde organisatie niet samenvallen. De EU AI Act kwalificeert AI in werving en werknemersbewaking als hoog-risico en verbiedt emotieherkenning in sollicitatiegesprekken, een eerste juridische verankering van werknemersrechten in het tijdperk van algoritmisch personeelsbeheer. Een rechtvaardige AI-transitie vereist dat de kosten en baten niet worden bepaald door marktdynamiek alleen, maar ook door regulering, sociale bescherming en werknemersparticipatie.

Slotwoord

Reflectie

> *Systemen die ontworpen worden om efficiënt te zijn, zijn dat ook. Wat ze efficiënt doen, is een andere vraag.*

Dat is de draad die door dit handboek loopt. AI-systemen zijn geen neutrale instrumenten die een wereld buiten zichzelf weerspiegelen. Ze coderen aannames, versterken patronen en nemen beslissingen, ook wanneer niemand ze zo heeft geprogrammeerd. Bias zit in de data. Surveillance zit in de infrastructuur. Ondoorzichtigheid zit in het businessmodel. Dat zijn geen technische gebreken die wachten op een patch. Het zijn structurele keuzes, die op elk punt in de ontwerp- en deploymentschain opnieuw gemaakt worden.

Achter die keuzes gaan belangen schuil. De organisaties die AI-systemen ontwikkelen en op grote schaal inzetten zijn niet willekeurig verdeeld over de wereld. Een handvol bedrijven en overheidsactoren bepaalt in grote mate welke systemen globaal worden uitgerold, op welke data ze getraind zijn en wat als normale output geldt. Dat is de structuur van een industrie die geografisch, cultureel en economisch geconcentreerd is. Wie buiten dat centrum valt, ontvangt technologie die niet voor hem ontworpen werd maar wel op hem wordt toegepast, en draagt de kosten van aannames die elders als vanzelfsprekend gelden.

Die concentratie heeft ook een rechtsstatelijke dimensie. De Europese pogingen om via regulering grip te krijgen op AI-systemen stuiten op een fundamentele spanning: de infrastructuur waarop die systemen draaien valt deels onder de rechtsmacht van derde landen. Digitale soevereiniteit benoemt een reëel machtsdeficit dat geen privacywet of AI-verordening volledig kan opvullen zolang de infrastructuurafhankelijkheid blijft bestaan. Europa reguleert wat actoren op Europees grondgebied met data mogen doen; het reguleert niet de juridische greep die een buitenlandse mogendheid op diezelfde data kan uitoefenen.

De vraag is niet of AI ethische vragen oproept. Die vraag is al beantwoord door de praktijk. De vraag is wie over die keuzes beslist, onder welke voorwaarden, en met welke verantwoording achteraf. Wetgeving zoals de EU AI Act, mechanismen zoals de GDPR, en concepten zoals explainability en meaningful human control zijn pogingen om die vraag institutioneel te beantwoorden. Ze zijn onvolledig, betwist en soms moeilijk afdwingbaar. Maar ze markeren een verschuiving: van een debat over wat AI *kan*, naar een debat over wat AI *mag*.

Regulering is niet hetzelfde als ethiek. Een systeem dat voldoet aan de letter van de EU AI Act kan nog steeds onrechtvaardig zijn. Compliance-kaders verplichten tot documentatie en risicoanalyse, maar schrijven geen waarden voor. De vraag welke schade aanvaardbaar is, wie die schade draagt en hoeveel transparantie werkelijk voldoende is, kan door wetgeving worden ingekaderd maar niet worden beantwoord. Dat antwoord vereist een oordeel, en dat oordeel is altijd politiek, altijd contextueel, altijd betwistbaar.

Voor burgers is AI-geletterdheid geen luxe maar een voorwaarde voor reëel burgerschap in een gedigitaliseerde samenleving. Wie niet begrijpt hoe een algoritme zijn kansen op een hypotheek of een sollicitatieprocedure beïnvloedt, kan dat algoritme niet betwisten. Het recht op uitleg is juridisch verankerd, maar een recht dat men niet begrijpt, kan men niet uitoefenen. De verantwoordelijkheid daarvoor ligt niet alleen bij de burger. Ze ligt bij de instellingen, de ontwerpers en de beleidsmakers die beslissen hoe toegankelijk die systemen worden gemaakt en hoe begrijpelijk de communicatie erover.

Die verschuiving vraagt om professionals die de technische logica begrijpen én de maatschappelijke gevolgen kunnen benoemen. Niet als sceptici die elke toepassing afwijzen, maar als mensen die weten welke vragen gesteld moeten worden voordat een systeem in gebruik gaat. Welke data liggen aan de basis? Wie draagt de schade bij een fout? Hoe wordt het systeem gecontroleerd wanneer de context verandert? Wie kan bezwaar aantekenen, en langs welke weg?

AI-ethiek is geen luxe voor wie genoeg tijd heeft om over principes na te denken. Ze is de voorwaarde waaronder technologie werkelijk dienstbaar kan zijn, aan individuen, aan instellingen, aan een samenleving die meer wil zijn dan de optelsom van haar optimalisatiefuncties.

APPENDIX

Referenties

Bronnen, per hoofdstuk geordend

Hoofdstuk 1. Introductie

ACADEMISCHE EN BELEIDSBRONNEN

01 Bengio, Y., Hinton, G., Yao, A., Russell, S. et al. (2024). "Managing extreme AI risks amid rapid progress". *Science*, 384(6698). https://doi.org/10.1126/science.adn0117

02 Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press / Picador.

03 Friedman, B. & Hendry, D.G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.

04 Noble, S.U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

05 Nussbaum, M. (2011). *Creating Capabilities: The Human Development Approach*. Harvard University Press.

06 Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

07 UNESCO (2021). *Recommendation on the Ethics of Artificial Intelligence*. Aangenomen door de 194 UNESCO-lidstaten, 23 november 2021. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

08 NIST (2023). *AI Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. https://www.nist.gov/itl/ai-risk-management-framework

09 Europese Unie (2024). Verordening (EU) 2024/1689, Artificial Intelligence Act. In werking getreden 1 augustus 2024. https://artificialintelligenceact.eu/

10 Westerstrand, S. (2024). "Reconstructing AI Ethics Principles: Rawlsian Ethics of Artificial Intelligence". *Science and Engineering Ethics*. https://link.springer.com/article/10.1007/s11948-024-00507-y

11 AlgorithmWatch (2019). "Industry defuses ethics guidelines for artificial intelligence" [over Metzingers ethics washing-kritiek]. https://algorithmwatch.org/en/industry-defuses-ethics-guidelines-for-artificial-intelligence/

12 Springer Nature / AI and Ethics (2024). "Digital ethicswashing: a systematic review and a process-perception-outcome framework". https://link.springer.com/article/10.1007/s43681-024-00430-9

ACTUELE BRONNEN (JOURNALISTIEK EN BELEID)

01 Computable.nl (2026). "Waarom digitale soevereiniteit in 2026 fundamenteel anders uitpakt". https://www.computable.nl/persberichten/waarom-digitale-soevereiniteit-in-2026-fundamenteel-anders-uitpakt/

02 SolidBE (2025). "Waarom Europa haar digitale soevereiniteit niet kan outsourcen aan Amerikaanse techgiganten". https://solidbe.nl/blog/artikelen/waarom-europa-haar-digitale-soevereiniteit-niet-kan-outsourcen-aan-amerikaanse-techgiganten/

03 DLA Piper (augustus 2025). "Latest wave of obligations under the EU AI Act take effect: Key considerations". https://www.dlapiper.com/en-us/insights/publications/2025/08/latest-wave-of-obligations-under-the-eu-ai-act-take-effect

04 Kennedys Law (2026). "The EU AI Act implementation timeline: understanding the next deadline for compliance". https://www.kennedyslaw.com/en/thought-leadership/article/2026/the-eu-ai-act-implementation-timeline-understanding-the-next-deadline-for-compliance/

Hoofdstuk 2. Bias

ACADEMISCHE BRONNEN

01 Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, 23 mei). "Machine Bias: There's software used across the country to predict future criminals. And it's biased against Blacks." *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

02 Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. https://fairmlbook.org/

03 Buolamwini, J., & Gebru, T. (2018). "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research*, 81, 77–91 (FAccT 2018). https://proceedings.mlr.press/v81/buolamwini18a.html

04 Kearns, M., & Roth, A. (2019). *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press.

05 Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). "Dissecting racial bias in an algorithm used to manage the health of populations." *Science*, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

06 Nizam Din, N., & Yasmin, M. (2023). "Ethics and discrimination in artificial intelligence-enabled recruitment practices." *Humanities and Social Sciences Communications*, 10(1). https://www.nature.com/articles/s41599-023-02079-x

07 Zhang, X., & Ding, Y. (2024). "Algorithmic discrimination in the credit domain: what do we know about it?" *AI & Society*, 38, 923–945. https://link.springer.com/article/10.1007/s00146-023-01676-3

08 Europese Unie (2024). Verordening (EU) 2024/1689, Artificial Intelligence Act, Bijlage III (high-risk toepassingen). https://artificialintelligenceact.eu/annex/3/

ACTUELE BRONNEN (JOURNALISTIEK EN RECHTSPRAAK)

01 The Washington Post (1 december 2025). "What to do if you fear AI is discriminating against you at work". https://www.washingtonpost.com/business/2025/12/01/ai-work-regulations-california/

02 The Guardian (11 augustus 2025). "AI tools used by English councils downplay women's health issues, study finds". https://www.theguardian.com/technology/2025/aug/11/ai-tools-used-by-english-councils-downplay-womens-health-issues-study-finds

03 The Guardian (5 november 2025). "Facebook's job ads algorithm is sexist, French equality watchdog rules". https://www.theguardian.com/world/2025/nov/05/facebook-job-ads-algorithm-is-sexist-french-equality-watchdog-rules

04 The Guardian (25 februari 2026). "Facial recognition error prompts police to arrest Asian man for burglary 100 miles away". https://www.theguardian.com/technology/2026/feb/25/facial-recognition-error-prompts-police-to-arrest-asian-man-for-burglary-100-miles-away

05 The Guardian (5 februari 2026). "'Orwellian': Sainsbury's staff using facial recognition tech eject innocent shopper". https://www.theguardian.com/technology/2026/feb/05/london-man-sainsburys-facial-recognition-facewatch

06 Financial Times (9 december 2025). "The perils of using AI when recruiting". https://www.ft.com/content/229983ee-c11f-44fb-8e61-2ac61d8d100a

07 Civil Rights Litigation Clearinghouse (16 mei 2025). "Order Granting Preliminary Collective Certification, Mobley v. Workday, Inc.". https://clearinghouse.net/doc/166151/

08 Europese Commissie, AI Act overzicht. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Hoofdstuk 3. Privacy en Surveillance

ACADEMISCHE EN THEORETISCHE BRONNEN

01 Nissenbaum, H. (2004). "Privacy as Contextual Integrity." *Washington Law Review*, 79(1), 119–158. https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10/

02 Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

03 Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

04 Büchi, M., Festic, N., & Latzer, M. (2022). "The Chilling Effects of Digital Dataveillance." *Big Data & Society*, 9. https://doi.org/10.1177/20539517211065368

05 Penney, J.W. (2022). "Understanding Chilling Effects." *Minnesota Law Review*, 106(3), 1451–1525. https://minnesotalawreview.org/article/understanding-chilling-effects/

06 Wang, X. et al. (2024). "Beyond Surveillance: Privacy, Ethics, and Regulations in Face Recognition Technology." *Frontiers in Big Data*, 7. https://doi.org/10.3389/fdata.2024.1337465

WETGEVING EN BELEID

01 Europese Unie (2016). Algemene Verordening Gegevensbescherming (GDPR), Verordening (EU) 2016/679, Art. 22 (geautomatiseerde besluitvorming). https://gdpr-info.eu/art-22-gdpr/

02 Europese Unie (2024). Artificial Intelligence Act, Verordening (EU) 2024/1689, Art. 5 (verboden AI-praktijken, waaronder realtime facial recognition). https://artificialintelligenceact.eu/article/5/

ACTUELE BRONNEN (JOURNALISTIEK)

01 The Guardian (10 december 2025). "UK police forces lobbied to use biased facial recognition technology." https://www.theguardian.com/technology/2025/dec/10/police-facial-recognition-technology-bias

02 The Guardian (25 februari 2026). "Facial recognition error prompts police to arrest Asian man for burglary 100 miles away." https://www.theguardian.com/technology/2026/feb/25/facial-recognition-error-prompts-police-to-arrest-asian-man-for-burglary-100-miles-away

03 The Guardian (6 april 2026). "'Creepy surveillance': why some cities are shutting down Flock cameras amid privacy concerns." https://www.theguardian.com/us-news/ng-interactive/2026/apr/06/flock-cameras-privacy-concerns

04 AP News (9 december 2025). "AI-powered police body cameras, once taboo, get tested on Canadian city's 'watch list' of faces." https://apnews.com/article/21f319ce806a0023f855eb69d928d31e

05 AP News (3 december 2025). "Takeaways from AP report on how Border Patrol monitors US drivers for 'suspicious' travel." https://apnews.com/article/48a6056d5661c676d33867afe4724464

Hoofdstuk 4. Explainability

ACADEMISCHE BRONNEN

01 Doshi-Velez, F. & Kim, B. (2017). *Towards a Rigorous Science of Interpretable Machine Learning*. arXiv:1702.08608.

02 Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press / Picador.

03 Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daume III, H. & Crawford, K. (2021). "Datasheets for datasets". *Communications of the ACM*, 64(12), 86–92.

04 Lipton, Z.C. (2018). "The mythos of model interpretability". *ACM Queue*, 16(3), 31–57.

05 Lundberg, S.M. & Lee, S.-I. (2017). "A unified approach to interpreting model predictions". In *Advances in Neural Information Processing Systems* (pp. 4765–4774). Curran Associates.

06 Mitchell, M. et al. (2019). "Model cards for model reporting". In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (pp. 220–229).

07 Ribeiro, M.T., Singh, S. & Guestrin, C. (2016). "'Why should I trust you?': Explaining the predictions of any classifier". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

08 Rudin, C. (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". *Nature Machine Intelligence*, 1(5), 206–215. https://www.nature.com/articles/s42256-019-0048-x

BELEID EN REGULERING

01 Europese Unie (2024). Verordening (EU) 2024/1689, EU AI Act. Artikel 13: Transparantie en informatieverstrekking aan deployers. https://artificialintelligenceact.eu/

02 Europese Unie (2016). Algemene verordening gegevensbescherming (GDPR). Artikel 22: Geautomatiseerde individuele besluitvorming.

03 Hof van Justitie van de Europese Unie (27 februari 2025). Zaak C-203/22, recht op uitleg bij geautomatiseerde besluitvorming; verwerkingsverantwoordelijken moeten de werkelijk toegepaste procedure en principes uitleggen, niet louter het algoritme beschrijven. https://www.insideprivacy.com/gdpr/cjeu-clarifies-gdpr-rights-on-automated-decision-making-and-trade-secrets/

04 TechPolicy.Press. "Understanding Right to Explanation and Automated Decision-Making in Europe's GDPR and AI Act". https://www.techpolicy.press/understanding-right-to-explanation-and-automated-decisionmaking-in-europes-gdpr-and-ai-act/

Hoofdstuk 5. AI Governance

ACADEMISCHE BRONNEN

01 Dafoe, A. (2018). *AI Governance: A Research Agenda*. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford.

02 Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People, an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707.

03 Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1), 99–120.

04 Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.

BELEID EN REGULERING

01 Europees Parlement en de Raad van de Europese Unie. (2024). *Verordening (EU) 2024/1689 betreffende artificiële intelligentie* (EU AI Act). Publicatieblad van de Europese Unie.

02 Europees Parlement en de Raad van de Europese Unie. (2022). *Verordening (EU) 2022/868 betreffende Europese datagovernance* (Data Governance Act). Publicatieblad van de Europese Unie.

03 National Institute of Standards and Technology (NIST). (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce.

04 OECD. (2019, herzien 2024). *OECD Principles on Artificial Intelligence*. Organisatie voor Economische Samenwerking en Ontwikkeling.

05 UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*. United Nations Educational, Scientific and Cultural Organization.

06 United States Congress. (2018). *Clarifying Lawful Overseas Use of Data Act* (CLOUD Act), 18 U.S.C. § 2713.

Hoofdstuk 6. GDPR en AI-Ethiek

ACADEMISCHE BRONNEN

01 Edwards, L., & Veale, M. (2017). Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for. *Duke Law & Technology Review*, 16(1), 18–84. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3052831

02 Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76–99. https://doi.org/10.1093/idpl/ipx005

WET- EN REGELGEVING

01 Verordening (EU) 2016/679 van het Europees Parlement en de Raad (AVG/GDPR), in het bijzonder Artikel 22 en Overweging 71.

02 Verordening (EU) 2024/1689 van het Europees Parlement en de Raad betreffende kunstmatige intelligentie (EU AI Act), in het bijzonder Artikel 13 (transparantie).

BELEIDSDOCUMENTEN EN INSTITUTIONELE BRONNEN

01 Belgische Gegevensbeschermingsautoriteit. (2022). *Artificial Intelligence Systems and the GDPR: A Data Protection Perspective*. https://www.autoriteprotectiondonnees.be/publications/artificial-intelligence-systems-and-the-gdpr---a-data-protection-perspective.pdf

Hoofdstuk 7. Aanvallen op AI-systemen

ACADEMISCHE BRONNEN

01 Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, 1807–1814. arXiv:1206.6389.

https://arxiv.org/abs/1206.6389

02 Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM Conference on Computer and Communications Security (CCS 2015)*, 1322–1333. https://doi.org/10.1145/2810103.2813677

03 Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*. arXiv:1412.6572. https://arxiv.org/abs/1412.6572

04 Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP 2017)*, 3–18. https://doi.org/10.1109/SP.2017.41

05 Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR 2014)*. arXiv:1312.6199. https://arxiv.org/abs/1312.6199

06 Tramer, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction APIs. In *Proceedings of the 25th USENIX Security Symposium*, 601–618. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer

RAPPORTEN EN BELEIDSDOCUMENTEN

01 European Union Agency for Cybersecurity (ENISA). (2024). *ENISA Threat Landscape 2024*. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024

02 National Institute of Standards and Technology (NIST). (2023). *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* (NIST AI 100-2). https://doi.org/10.6028/NIST.AI.100-2e2023

Hoofdstuk 8. Mensenrechten

ACADEMISCHE BRONNEN

01 Alston, P. (2019). *Report of the Special Rapporteur on extreme poverty and human rights*. United Nations General Assembly, document A/74/493. https://undocs.org/A/74/493

02 Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

03 Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, 13(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

04 Noble, S.U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

05 Nussbaum, M.C. (2011). *Creating Capabilities: The Human Development Approach*. Harvard University Press.

06 Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76–99. https://doi.org/10.1093/idpl/ipx005

07 Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

BELEIDSDOCUMENTEN EN INSTITUTIONELE BRONNEN

01 UNESCO (2021). *Recommendation on the Ethics of Artificial Intelligence*. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

02 Verordening (EU) 2024/1689 van het Europees Parlement en de Raad betreffende kunstmatige intelligentie (EU AI Act), in het bijzonder artikel 6 en Bijlage III (hoog-risicosystemen). https://artificialintelligenceact.eu/

Hoofdstuk 9. Autonome Systemen

ACADEMISCHE BRONNEN

01 Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, 23 mei). Machine Bias: There's software used across the country to predict future criminals. And it's biased against Blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

02 Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, 563, 59–64. https://doi.org/10.1038/s41586-018-0637-6

03 Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

04 Human Rights Watch. (2012). *Losing Humanity: The Case against Killer Robots*. https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots

05 Lundberg, S.M., & Lee, S.I. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems* (NeurIPS 2017), 30. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

06 Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

07 Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

08 Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

09 Topol, E.J. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books.

BELEIDSDOCUMENTEN EN INSTITUTIONELE BRONNEN

01 Europese Commissie. (2020). *White Paper on Artificial Intelligence: A European approach to excellence and trust*. COM(2020) 65 final. https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

02 United Nations General Assembly. (2024). *Resolution on Lethal Autonomous Weapons Systems* (A/RES/79/67). Aangenomen op 23 december 2024 met 166 stemmen voor.

Hoofdstuk 10. De Toekomst van Werk

ACADEMISCHE BRONNEN

01 Acemoglu, D. (2024). *The Simple Macroeconomics of AI*. NBER Working Paper 32487. https://doi.org/10.3386/w32487

02 Acemoglu, D., & Restrepo, P. (2020). Robots and Jobs: Evidence from US Labor Markets. *Journal of Political Economy*, 128(6), 2188–2244. https://doi.org/10.1086/705716

03 Berg, J., Furrer, M., Harmon, E., Rani, U., & Six Silberman, M. (2018). *Digital Labour Platforms and the Future of Work: Towards Decent Work in the Online World*. International Labour Organization.

04 Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384, 1306. https://doi.org/10.1126/science.adj0998

05 Frey, C.B., & Osborne, M.A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

06 Nussbaum, M. (2011). *Creating Capabilities: The Human Development Approach*. Harvard University Press.

07 O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

08 Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

09 Weil, D. (2014). *The Fissured Workplace: Why Work Became So Bad for So Many and What Can Be Done to Improve It*. Harvard University Press.

BELEIDSDOCUMENTEN EN INSTITUTIONELE BRONNEN

01 Europees Parlement en de Raad van de Europese Unie. (2024). *Verordening (EU) 2024/1689 betreffende artificiële intelligentie (AI-verordening)*, Bijlage III. Publicatieblad van de Europese Unie.

02 International Labour Organization. (2025). *Generative AI and jobs: A 2025 update*. ILO. https://www.ilo.org/publications/generative-ai-and-jobs-2025-update

03 World Economic Forum. (2025). *Future of Jobs Report 2025*. https://www.weforum.org/publications/the-future-of-jobs-report-2025/