Ontsluit data van big tech via statistische wetgeving

Tech-giganten zoals Google, Meta, Microsoft en Amazon hebben een goudmijn aan data. Deze data en de kennis die ze op basis van die data ontwikkelen, is niet publiek beschikbaar. Hierdoor dreigt een kenniskloof te ontstaan tussen tech-giganten en de samenleving. Wat kunnen we doen om een 'dystopische kennis-kloof' te vermijden?

IN HET KORT

- De data van big tech openbaar maken is onmogelijk door privacy-problemen.
- De data beschikbaar maken voor statistiek en wetenschap is wel mogelijk.
- Het beschikbaar maken van de data kan relatief eenvoudig via de bestaande statistische wetgeving.

HENK VAN TUINEN
Voormalig plaatsvervangend directeur-generaal van het Centraal Bureau voor de Statistiek

Ondernemingen zoals Google, Meta, Microsoft of Amazon kennen jou misschien wel beter dan je levenspartner jou kent. Door alle sporen die je nalaat op internet vast te leggen in hun databestanden, en deze met hun superieure kunstmatige intelligentie (AI) grondig te analyseren, creëren ze van jou een uitermate gedetailleerd profiel – waarin je gedrag en je karaktereigenschappen zijn vastgelegd (Zuboff, 2019).

Door honderden miljoenen van die gedetailleerde profielen beschikbaar te maken voor *targeted advertising* verdienden de tech-giganten zo veel geld dat ze binnen vijftien jaar de wereldwijde omzet van de tv-reclame hebben overtroffen (Frederik en Martijn, 2019) en uitgroeiden tot de waardevolste multinationals. Het spreekt dan ook vanzelf dat ze die goudmijn graag voor zichzelf houden.

De gemiddelde econoom zal zich vanwege de marktmacht echter zorgen maken over inefficiëntie. Er zijn dan ook diverse economisch beargumenteerde voorstellen gedaan om de macht van de tech-giganten in te perken: regulering, opsplitsing (Ovide, 2021), herverdeling van monopoliewinsten (Romer, 2019), en het laten betalen voor de data die men oogst op internet (Posner en Weyl, 2018).

Wie breder kijkt, ziet ook andere problemen opdoemen. Aan het eind van zijn bestseller *Homo Deus* voorspelt de historicus Yuval Noah Harari het uitsterven van de homo sapiens (Harari, 2017). De mens zal worden vervangen door algoritmes. Hij voorziet dat wij behandeld zullen worden, zoals we nu onze huisdieren behandelen. In haar boek ontwikkelt Zuboff (2019) een theorie die, zonder dat zij naar Harari verwijst, de perfecte grondslag biedt voor Harari's dystopie. Zuboff stelt vast dat de tech-giganten ons gedrag steeds preciezer kunnen voorspellen en dat ze het bij gelegenheid ook manipuleren. In haar analyse zal dit leiden tot "a division of learning": tech-giganten weten alles van ons – en monopoliseren die superieure kennis. Wij worden tweederangsburgers met inferieure kennis.

Filosofisch, maatschappelijk en politiek bezien is er dus sprake van een doemscenario. De recente paniek rond de stormachtige ontwikkeling van AI, naar aanleiding van de hype rond ChatGPT, illustreert dat dit doemdenken zich snel kan verbreiden (Harari et al., 2023; Metz en Schmidt, 2023).

In april jongstleden kon de *ESB*-lezer kennisnemen van een voorstel om de data van tech-giganten openbaar te maken (De Ridder, 2023). Het voorstel van De Ridder kent echter praktische bezwaren. In dit artikel licht ik deze bezwaren toe, waarna ik zelf een nieuw voorstel aandraag dat wel praktisch haalbaar is.

Openbaar maken data problematisch

Het plan om data van big tech openbaar te maken stuit op tenminste twee problemen. Allereerst zullen de tech-giganten er op wijzen dat ze veel hebben geïnvesteerd om de goudmijn aan data aan te leggen, dat ze het recht hebben om die te exploiteren, en dat derden dat recht niet zomaar hebben. En zeker niet zonder de giganten daarvoor ruimschoots schadeloos te stellen.

Het tweede probleem betreft de beveiliging van persoonsgegevens. Big tech zal benadrukken dat het bij hun data gaat om persoonsgegevens die niet voldoende beveiligd kunnen worden. Dat zal in de praktijk een uiterst krachtig argument opleveren tegen het openbaar maken van die rijke profielen.

Bij het beveiligen van persoonsgegevens gaat het namelijk om veel meer dan de-identificeren. Een instelling die over persoonsgegevens beschikt, moet ervoor zorgen dat derden nooit de identiteit kunnen afleiden van degene op wie de gegevens betrekking hebben. Als het gaat om een rijke verzameling gedetailleerde gegevens, is de kans altijd aanwezig dat de identiteit kan worden vastgesteld, ook als alle directe identificatoren verwijderd zijn. Hoe gedetailleerder het dataprofiel (oftewel de per individu opgeslagen reeks gegevens), hoe groter de kans. Dat laat zich al eenvoudig inzien zodra je bedenkt dat bij een extreem gedetailleerd dataprofiel elk individu uniek is. Bijvoorbeeld: als in het dataprofiel detailgegevens zijn samengebracht over levensloop, beroepsuitoefening en online-koopgedrag,

wie is er dan niet uniek? Bedenk daarbij dat big tech veelal over nog veel rijkere dataprofielen beschikt. Bedenk ook dat gegadigden voor de data van big tech, die zelf over persoonsgegevens beschikken, eenvoudig door bestandskoppeling overeenkomstige records kunnen identificeren. Je apotheek heeft al genoeg aan je geboortedatum.

Big tech zal, indien ze worden gedwongen tot openbaar maken van de data, natuurlijk tot het uiterste gaan bij de bescherming van persoonsgegevens. De Ridder merkt in zijn voorstel op: "Uiteraard worden de data ontdaan van persoonsgegevens en zijn bedrijfsgeheimen uitgesloten." Dat betekent dat dit slechts tot een uitermate beperkte, inferieure deelverzameling van hun profielen zal leiden, met verregaand 'ingedikte' records. Het resultaat is een sterk gereduceerde dataverzameling, waarmee de economische inefficiëntie niet wezenlijk wordt bestreden en de kenniskloof evenmin.

Een praktisch voorstel

Er is een methode waarmee de goudmijn zo beschikbaar kan worden gesteld dat het doemscenario van de kenniskloof kan worden vermeden. Daar is waarschijnlijk zelfs geen nieuwe wetgeving voor nodig, omdat de bestaande statistische wetgeving al een goede basis levert. Die methode stelde ik eerder voor in een artikel met een breder onderwerp (Van Tuinen, 2021). Voor achtergronden en details verwijs ik graag naar dat artikel.

In westerse landen bestaat er statistische wetgeving die ondernemingen verplicht om gegevens over hun productie en exploitatie te leveren aan het nationale statistische instituut – in Nederland is dat aan het CBS (2023a). Het kan daarbij gaan om uiterst gedetailleerde gegevens over hun activiteiten. Die wettelijke verplichting is veelal in de eerste helft van de vorige eeuw ingesteld met als doel dat er nauwkeurige statistieken over het economisch proces konden worden samengesteld door de statistische instituten.

De statistische instituten mogen de aldus verzamelde gegevens niet zo publiceren dat daaruit informatie over individuele ondernemingen of andere entiteiten kan worden afgeleid. Ook mogen zij geen identificeerbare gegevens doorgeven aan de overheid of de rechterlijke macht. Maar zij kunnen wel over alle identificerende details beschikken bij hun statistische analyses.

Gezien het grote belang van het overbruggen van de kenniskloof zou de goudmijn kunnen worden opgeëist op basis van de statistische wetgeving. Ondernemingen kunnen zich daar niet tegen verzetten op grond van bedrijfsgeheimen of bescherming van persoonsgegevens. Dat is redelijk, omdat de statistische instituten geen individuele gegevens onthullen. Statistiek gaat over aggregaten en verbanden, niet over individuele entiteiten.

Door toepassing van de statistische wetgeving kunnen de data van big tech beschikbaar komen voor statistisch en wetenschappelijk onderzoek door alle bonafide wetenschappelijke onderzoeksinstellingen, ten behoeve van iedereen in onze samenleving. Zo kan de kenniskloof effectief worden overbrugd. En zo kan dus het doemscenario dat onze menselijkheid bedreigt, worden vermeden.

Goed uitvoerbaar

De ervaring met de statistische wetgeving leert dat bedrijven natuurlijk niet altijd blij zijn met de opgelegde informatieplicht. De tech-giganten zullen dus wel weer geducht gaan lobbyen als het voorstel dreigt te worden uitgevoerd. Maar ze hebben daarbij weinig argumenten beschikbaar. Juist omdat de te leveren gegevens niet openbaar worden, kunnen zij zich niet beroepen op hun bedrijfsbelang. Wanneer de informatieplicht ook geldt voor hun concurrenten, blijft er nauwelijks een argument over. Vroeger kregen bedrijven wel eens gehoor bij hun bezwaar tegen de 'administratieve last' van de informatieverstrekking – maar dat zal in het geval van tech-giganten met de meest geavanceerde ICT niet gauw serieus genomen hoeven te worden. Dat de tech-giganten zich bij levering aan statistische instituten geen zorgen hoeven te maken over de beveiliging van hun persoonsgegevens maakt de data-verstrekking voor hen eenvoudig uitvoerbaar.

De publieke opinie zal zich niet gauw aan hun zijde scharen, want hun ongebreidelde gebruik van de data over ons gedrag was nooit onze bedoeling toen wij die data genereerden door onze verslaving aan hun apps. De tech-giganten komen de laatste jaren niet voor niets regelmatig onder vuur te liggen.

Maar kunnen de nationale statistische instituten al die *big data* van big tech wel aan? Ze hebben er al veel ervaring mee opgedaan. Zo heeft het Nederlandse CBS geleerd om zeer grote databestanden te koppelen en te analyseren, inclusief big data (CBS, 2023b). En het heeft geleerd daarbij intensief samen te werken met externe onderzoeksinstellingen (CBS, 2023c). Misschien zijn niet alle westerse statistische instituten al even ver als ons CBS, maar dat mag geen reden zijn om bijvoorbeeld de Europa-brede uitvoering van het voorstel af te wijzen. Veeleer is dat een reden om extra te investeren in de onafhankelijkheid en professionaliteit van nationale statistische instituten.

De statistische instituten zullen moeten kunnen investeren in hun ICT en analyse-capaciteit om mijn voorstel uit te kunnen voeren. Wellicht geldt dat ook voor wetenschappelijke onderzoeksinstellingen die met de statistische instituten willen samenwerken. Maar dan is er sprake van een *blessing in disguise*: de ontginning van de goudmijn zal onze kennis een *boost* gaan geven.

Het investeren in onafhankelijkheid en professionaliteit van nationale statistische instituten is extra welkom in een tijd waarin misinformatie, polarisatie en verscherping van de internationale tegenstellingen leiden tot een algemene vermindering van het vertrouwen in de wetenschap – met alle gevolgen van dien voor economie, democratie en veiligheid. Het is belangrijk om te investeren in het vertrouwen in statistiek en wetenschap als de pijlers waarop onze informatiesamenleving kan blijven rusten.

Terwijl er internationaal moeizaam geworsteld wordt met het op gang brengen van een noodzakelijke regulering van de informatiestromen die via de tech-giganten lopen, kan met dit voorstel het overbruggen van de kenniskloof verrassend eenvoudig worden gerealiseerd. En zo wordt een enorm onheil voorkomen.

Geen panacee

Het beschikbaar maken van de data van big tech via statistische wetgeving is efficiënter dan het voorstel van De Ridder tot openbaarmaking, want dat zal na moeizame wetgeving en eindeloos juridisch getouwtrek met tech-giganten hooguit leiden tot hergebruik van een zeer beperkte en inferieure deelverzameling van de data. Mijn voorstel is betrekkelijk eenvoudig realiseerbaar op basis van bestaande wetgeving.

Een nadeel van mijn voorstel is dat de data alleen beschikbaar worden voor statistisch en wetenschappelijk onderzoek, maar niet voor andere doeleinden, zoals commercieel gebruik.

Als we breder kijken dan de economische efficiëntie, is het voorstel geen panacee. Het maakt geen einde aan de surveillance door de tech-giganten. Alle mogelijkheden tot manipulatie van ons gedrag, door de giganten zelf of door hun klanten – waaronder vooral invasieve reclamemakers (Van Tuinen, 2021) – blijven in stand. Deze problemen zullen moeten worden opgelost met vergaande regulering. Hetzelfde geldt voor het probleem van de misinformatie die onze samenleving en democratie bedreigt. Maar die regulering zal niet zomaar een oplossing bieden voor het probleem van de dystopische kenniskloof dat centraal staat in dit artikel.

Literatuur

CBS (2023a) *Wet- & regelgeving*. CBS Informatie.

CBS (2023b) *Center for Big Data Statistics*. CBS Informatie.

CBS (2023c) *Microdata: Zelf onderzoek doen*. CBS informatie.

Frederik, J. en M. Martijn (2019) *The new dot com bubble is here: It's called online advertising*. Artikel op thecorrespondent.com, 6 november.

Harari, Y.N. (2017) *Homo Deus: Een kleine geschiedenis van de toekomst*. Amsterdam: Thomas Rap.

Harari, Y., T. Harris en A. Raskin (2023) You can have the blue pill or the red pill, and we're out of blue pills. *The New York Times*, 24 maart.

Metz, C. en G. Schmidt (2023) Elon Musk and others call for pause on A.I., Citing 'Profound Risks to Society. *The New York Times*, 29 maart.

Ovide, S. (2021) The state house versus big tech. State and local governments are looking to assert more control over tech companies. *The New York Times*, 16 februari.

Posner, E.A. en E.G. Weyl (2018) *Radical markets: Uprooting capitalism and democracy for a just society*. Princeton: Princeton University Press.

Ridder, W.P. de (2023) *Boekbespreking: Maak de data van big tech openbaar*. Blog op esb.nu, 26 maart.

Romer, P. (2019) A tax that could fix big tech. *The New York Times*, 6 mei.

Tuinen, H.K. van (2021) A political economy of reorientation: New theory and policy for the recovery. *Central European Review of Economics and Management*, 5(2), 7–53.

Zuboff, S. (2019) *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Londen: Profile Books.