Waarheidsvinding in digitale tijden

DE MORELE VERANTWOORDELIJKHEID VAN SOCIALE MEDIA PLATFORMEN OMTRENT INFORMATION DISORDERS

Kalliopi Zoï Argyraki
2015070 | U394674
K.z.argyraki@tilburguniversity.edu

Bachelor scriptie filosofie | 701THE-B-10
Supervisor: dr. B. Engelen
Tweede lezer: dr. F.A.I. Buekens
Tilburg School of Humanities and Digital Sciences
Tilburg, juni 2021

Abstract

Het jaar 2020 heeft pijnlijk blootgelegd hoe mis- en desinformatie, die zich snel verspreid via sociale media platformen, een enorme impact kan hebben op onze democratie. Deze scriptie onderzoekt wat de morele verantwoordelijkheid is van sociale media platformen betreffende het verspreiden, screenen en/of censureren van *information disorders*, in het licht van vrijheid van meningsuiting. Ik beargumenteer dat het business model van de bedrijven achter de digitale platformen in combinatie met onze psychologische neigingen, waar deze bedrijven op inspelen, een broeiplaats vormen voor *information disorders*. Vervolgens beargumenteer ik dat sociale media platformen zoals Facebook, Twitter en YouTube een morele verantwoordelijkheid hebben om actief *information disorders* te bestrijden op basis van het principe 'waarheidsvinding'. Aan de hand van Mill en Habermas laat ik zien wat de meerwaarde van waarheidsvinding is in een democratie. Daarnaast beargumenteer ik aan de hand van Mill dat vrijheid van meningsuiting een middel is tot het doel waarheidsvinding en analyseer ik de implicaties hiervan voor de rol van *information disorders*. Ten slotte, concludeer ik dat het censureren van *information disorders* door de platformen zelf niet de oplossing is: willen we het probleem van *information disorders* aanpakken, dan moeten we de bron zelf (de sociale media platformen) kritisch aanpakken.

Key words: Sociale media platformen, information disorders, morele verantwoordelijkheid, waarheidsvinding, democratie, vrijheid van meningsuiting.
Words: 12.837

Voorwoord

In de vierde klas van de Havo schreef ik voor het vak Nederlands een betoog over waarom sociale media platformen ons als mensheid zo veel hebben gebracht. Mijn tegenargument op de objectie ‘dat er toch ook wel veel mis ging op het internet’ was het volgende: “sociale media platformen zijn in zekere zin net als vuur, het kan ons veel innovatie brengen in de toekomst maar je moet wel eerst een paar keer je vingers verbranden”. Nu een aantal jaar later schrijf ik mijn scriptie over hoe we onze vingers toch wel flink aan het verbranden zijn, al dan niet een vuurtje hebben gestart op een plek waar we het lastig kunnen blussen. Het moge duidelijk zijn dat ik het onderwerp ‘sociale media’ al een lange tijd fascinerend vind. Vooral het afgelopen jaar heb ik met fascinatie - maar ook angst - toegekeken hoe (onjuiste) informatie via digitale platformen een machtige positie binnen onze samenleving heeft verkregen. De bachelor filosofie heeft mij geleerd om deze fascinatie en angst om te zetten in een kritische analyse. Of we die machtige positie inderdaad enkel voor innovatie gaan gebruiken, gaat de tijd ons vanzelf leren.

Via deze weg zou ik graag mijn scriptiebegeleider Bart Engelen willen bedanken. Naast dat ik de begeleiding als een heel fijn proces heb ervaren, heb ik er enorm veel van geleerd. De uitgebreide, heldere feedbacksessies en de vele suggesties hebben mij goed door dit leuke maar soms ook lastige proces heen gesleept. Daarnaast wil ik graag mijn docenten van het departement filosofie bedanken die mij de afgelopen drie jaar hebben begeleid en geïnspireerd. Tot slot zou ik graag mijn vriend Yannick, mijn ouders en grootouders willen bedanken voor alle hulp: bedankt voor alle onvoorwaardelijke steun in alles, altijd.

Inhoudsopgave

1. Introductie

Op 6 januari 2021 bestormden honderden mensen het Capitool in Washington D.C.. Enkele uren daarvoor sprak de toenmalige Amerikaanse president Donald Trump een grote groep supporters toe met de oproep om 'Amerika te redden'. Amerika diende gered te worden omdat de verkiezingen volgens Trump 'gestolen' zouden zijn door de Democraten. De uitslag van de verkiezingen zou namelijk, wat overigens nooit bewezen is tot nu toe, gesaboteerd zijn. Gedurende de bestorming kregen de tweets van de president een waarschuwingslabel mee van Twitter, dat stelde dat de claims ongefundeerd waren, maar na een tijd werd zijn gehele account door het media platform offline gehaald. Ook platformen zoals Facebook en Instagram blokkeerden vervolgens Trumps sociale media-accounts. In een verklaring op Facebook liet Mark Zuckerberg, de CEO van Facebook en Instagram, weten dat deze ingrepen verdere geweldsescalatie moesten voorkomen (Zuckerberg, 2021).

Het is niet de eerste keer dat sociale media platformen ingrijpen door accounts te verwijderen. Zo heeft Facebook de afgelopen maanden meer dan 10.000 groepen en pagina's offline gehaald die werden geassocieerd met QAnon, een rechtsextremistische complottheorie (Wong, 2020). Ook in Nederland kennen we ondertussen een aantal voorbeelden van accounts die offline zijn gehaald door sociale media platformen. Zo werden de accounts van politicus Wybren van Haga en opiniepeiler Maurice de Hond verwijderd door LinkedIn en verwijderde het platform YouTube het account van zanger Lange Frans (NOS, 2020). De redenen voor het verwijderen van deze accounts kwamen steeds op hetzelfde neer: er werd desinformatie verspreid. De verspreiding van mis- en desinformatie via sociale media platformen zoals Facebook, Twitter en YouTube lijkt een steeds groter probleem te worden. "We're not just fighting an epidemic; we're fighting an infodemic", zei de directeur-generaal van de Wereldgezondheidsorganisatie tijdens de Munich Security Conference (Ghebreyesus, 2020). Hoe is deze 'infodemie' ontstaan? En, is het censureren van mis- en desinformatie de oplossing?

Deze scriptie buigt zich over dit thema en onderzoekt de vraag: wat is de morele verantwoordelijkheid van sociale media platformen betreffende het verspreiden, screenen en/of censureren van *information disorders,* in het licht van vrijheid van meningsuiting? Om de vraag te kunnen onderzoeken bespreken we eerst in hoofdstuk twee de termen rondom informatie. Wat betekent informatie in ons digitale tijdperk en waar kan het dusdanig fout gaan met informatie dat het schadelijk wordt? Deze termen vatten we samen onder de parapluterm *information disorder*. Vervolgens beargumenteer ik in hoofdstuk drie waarom de huidige

sociale media platformen een broeiplaats zijn voor *information disorders*. Het business model en het design achter de sociale media platformen in combinatie met onze psychologische neigingen waar de platformen op inspelen zorgt er voor dat *information disorders* zich makkelijk verspreiden. In hoofdstuk vier beargumenteer ik waarom sociale media platformen als morele actoren gezien moeten worden en op welk ethisch principe (waarheidsvinding) we ze moreel verantwoordelijk moeten houden. Ten slotte onderzoek ik in hoofdstuk vijf of er een spanning is tussen het principe 'waarheidsvinding' en het principe 'vrijheid van meningsuiting'. Men spreekt namelijk weleens van een schending van 'vrijheid van meningsuiting' wanneer informatie wordt gecensureerd op de digitale platformen. Aan de hand van John Stuart Mills bekende boek 'On Liberty' laat ik zien dat 'vrijheid van meningsuiting' een middel is tot het streven naar het doel 'waarheidsvinding'. Daarnaast beargumenteer ik aan de hand van Mill dat *information disorders* doorgaans niet onder vrijheid van meningsuiting vallen. Echter is het enkel censureren van *information disorders* door sociale media platformen niet genoeg. Ik concludeer op basis van de voorgaande hoofdstukken dat we als liberale, democratische samenleving verandering moeten eisen van de digitale platformen: of ze moeten zich aan bepaalde (design) principes gaan houden of ze stappen uit de nieuwssector en profileren zich weer als louter entertainment.

2. Information Disorders

We worden er haast mee overspoeld; informatie is overal om ons heen. Bewust maar vaak ook onbewust consumeren wij dagelijks talloos veel informatie. Door de enorme impuls die digitale media heeft gekregen in de afgelopen twee decennia consumeren we informatie op ieder moment van de dag via onze laptop of telefoon. Zo spendeerden volwassen Amerikanen in 2018 gemiddeld 6.3 uren per dag aan digitale media, terwijl we in 2008 nog maar konden spreken over 2.7 uren per dag (Meeker, 2019: 42). We worden meer dan ooit blootgesteld aan informatie, maar wat definieert het concept 'informatie' precies? Bovendien wanneer mogen we spreken van de veelgebruikte termen zoals 'desinformatie' en 'nepnieuws'? Het is cruciaal om eerst grip te krijgen op deze concepten voordat we ze in de volgende hoofdstukken toepassen. In paragraaf 2.1 definiëren we daarom eerst het concept 'informatie' en in paragraaf 2.2 definiëren we termen zoals 'desinformatie', 'misinformatie' en 'nepnieuws'. Uiteindelijk gebruik ik in andere hoofdstukken de parapluterm *information disorders* omdat ik op deze wijze makkelijk *alle* vormen van feitelijk onjuiste en/of misleidende informatie kan duiden.

2.1. Wat is informatie?

In ons dagelijks taalgebruik komt het woord ‘informatie’ vaak voor. Maar het definiëren van ‘informatie’ is nog niet zo makkelijk: zo zijn er diverse definities te vinden die zeer uiteen kunnen lopen. Zo schreef Wiener (1954) bijvoorbeeld dat ‘informatie’ de benaming is voor de inhoud van dat wat wordt uitgewisseld met de buitenwereld terwijl wij ons daarop aanpassen (17). Hij focuste zich met deze definitie meer op het proces achter informatie. Informatie kan ook gedefinieerd worden als een belangrijk epistemisch[1] goed dat in staat is om kennis te genereren (Dretske, 1983: 57). Informatie heeft daarmee dus waarde omdat het een waardevolle bijdrage levert aan het proces van kennisontwikkeling. Kennis kunnen we dan definiëren als een door informatie veroorzaakte overtuiging (Dretske, 1983: 58).

Wanneer je de betekenis van ‘informatie’ opzoekt in het woordenboek, dan krijg je de volgende definitie: “inlichtingen, gegevens waardoor je meer over iets te weten komt” (Van Dale, 2021). We nemen daarbij aan dat de proposities die de informatie inhoudelijk karakteriseren, vervolgens dan ook de wereld representeren, vooral als wij die proposities als waar achten (Buekens, 2018: 16). In andere woorden, informatie is dan iets dat een deel van de wereld representeert als ‘op een bepaalde manier zijn’. Een voorbeeld dat vaak wordt gebruikt, is de zin: “de kat ligt op de mat”. Deze zin representeert dat de kat echt op de mat ligt. Dit kun je, bij wijze van spreken, controleren door even te kijken bij de kat of door een foto te bekijken van de kat. Op deze manier zien we als kenmerk van informatie dat het semantische inhoud heeft (Floridi, 2011: 80).

Is informatie dan altijd ‘waar’? Volgens sommigen omvat semantische informatie per definitie waarheid (Floridi, 2011: 82; Dretske, 1983: 57). Zinsconstructies zoals ‘de juiste informatie’ en ‘de ware informatie’ zouden om die reden simpelweg pleonasmen zijn. Echter is niet iedereen het met deze opvatting eens: zo is er informatie dat waar, onwaar of allebei kan zijn (Fetzer, 2004). Ik ga niet verder in op de discussie wat ‘waarheid’ precies is. De vraag naar waarheid valt namelijk buiten de scope van mijn scriptie: deze scriptie heeft meer een normatieve focus dan een epistemologische focus. Echter neem ik wel de positie in dat informatie zonder waarheid kan bestaan en dat we dit kunnen definiëren als ‘feitelijk onjuiste informatie’ en ‘foutieve informatie’.

[1] Epistemologie heeft, in brede zin, te maken met cognitieve succes en cognitieve mislukking. Het houdt zich vooral bezig met de vraag wat kennis is en hoe we kennis genereren (Steup & Neta, 2020).

We kunnen er in ieder geval vanuit gaan dat ‘informatie’ - in de context van mijn scriptie - het beste begrepen kan worden als een fenomeen dat betekenis en waarde heeft: het brengt ons dichter bij kennis. Onderweg naar deze kennis, kan er echter veel epistemisch fout gaan.

2.2. Informatie zonder waarheid?

Informatie heeft waarde (het kan leiden tot kennis) en wanneer iets waarde heeft, kan het misbruikt worden. Een vorm van het misbruiken van informatie is het manipuleren van informatie. En dat is precies wat er aan de hand is wanneer we spreken over termen zoals ‘misinformatie’, ‘desinformatie’ en ‘nepnieuws’.

Misleidende en feitelijk onjuiste informatie kan dus onderverdeeld worden in verschillende concepten die overlap hebben met elkaar, maar wel vaak een andere definitie kennen. Het eerste wat we kunnen onderscheiden zijn de termen ‘misinformatie’ en ‘desinformatie’. Zo spreken we van ‘misinformatie’ wanneer het gaat om (deels) onjuiste informatie dat een misleidend resultaat kan opleveren (Lazer et al., 2018: 1094.) Hierbij is er geen bewuste intentie om foutieve informatie te verspreiden vanuit de verspreider. Denk hierbij, ter illustratie, aan een persoon die jou de weg verkeerd uitlegt omdat diegene per ongeluk “de tweede weg links” zegt in plaats van “de derde weg links”. In deze situatie ben je op het verkeerde spoor gezet maar er was geen bewuste intentie om je de verkeerde informatie te verschaffen. Bij ‘desinformatie’ spreken we daarentegen over onjuiste informatie die wel wordt verspreid met de intentie om mensen te misleiden (Ibid.). Hierbij valt te denken aan hoe verschillende Russische nepaccounts het stemgedrag van Europese burgers hebben proberen te beïnvloeden door feitelijk onjuiste informatie te verspreiden (Musch, 2019). Uit onderzoek bleek dat ongeveer de helft van de Europese kiezers deze feitelijk onjuiste informatie heeft ontvangen (Ibid.).

Een ander concept dat vaak gebruikt wordt in de media is ‘nepnieuws’. Het gebruik van de term ‘nepnieuws’ is erg populair geworden de laatste jaren. Nepnieuws kan gedefinieerd worden als mis- of desinformatie verhuld als nieuws, maar dan zonder de normen en processen die nodig zijn om accuraat nieuws te produceren (Lazer et al., 2018: 1094). Een vorm van nepnieuws zijn bijvoorbeeld de artikelen van ‘De Speld’, een satirisch platform dat artikelen publiceert alsof het ‘echt’ nieuws is. Ik beschouw ‘De Speld’ als een ‘onschuldige’ variant omdat het algemeen bekend is dat de artikelen fungeren als entertainment en altijd satirisch van

aard zijn[2]. Een vorm van kwalijker nepnieuws zijn de talloze berichten vermomd als 'echt' nieuws over bijvoorbeeld het coronavirus. Zo werden er berichten vanuit troll-accounts verspreid die pretendeerden dat het virus wordt veroorzaakt door 5G-netwerken of door miljardairs zoals Bill Gates en George Soros (Visser, 2020). De troll-accounts die op Twitter werden geanalyseerd hadden veelal rechts-extremistische motieven (Ibid.). Bij 'De Speld' spreken we daarom eerder over misinformatie verhuld als nieuws (hier is de intentie om grappig te zijn en dit is bekend bij de volgers) terwijl we bij het voorgaande voorbeeld over corona nepnieuws spreken over desinformatie verhuld als nieuws (er is namelijk wel sprake van een verkeerde intentie, namelijk: onrust zaaien).

Een verzamelterm die gebruikt wordt om alle bovenstaande concepten onder één concept vast te leggen is 'information disorder' (Lazer et al., 2018; Wardle, 2017). Om het breed maar toch overzichtelijk te houden verwijs ik vanaf nu naar *information disorders* wanneer ik het over 'desinformatie', 'nepnieuws' en in sommige gevallen 'misinformatie' heb.

Verschillende termen die onder *information disorder* vallen zijn relatief nieuwe termen, terwijl het inhoudelijk gaat om situaties die helemaal niet zo nieuw zijn. Zo stelt historicus Robert Darnton (2017) dat vormen van mis- en desinformatie in vrijwel iedere periode in de geschiedenis terug te vinden zijn. In één van zijn voorbeelden beschrijft hij hoe schrijver Pietro Aretino de pauselijke verkiezingen in 1522 probeerde te manipuleren door gemene verhalen te schrijven over alle kandidaten, behalve over zijn opdrachtgever. Mensen misleiden met verkeerde informatie is dus ook geen nieuw fenomeen. Door digitale technologieën gaat het verspreiden van *information disorders* echter in een snel tempo met als grootste verschil, ten opzichte van de voorbeelden van Darnton, dat het over de hele wereld verspreid wordt. In het volgende hoofdstuk wordt beschreven hoe en waarom *information disorders* zich zo makkelijk ontwikkelen en verspreiden op sociale media platformen.

3. Hoe information disorders verspreiden op sociale media platformen

Sociale media platformen zoals Facebook, Twitter, YouTube, Instagram en TikTok zijn een belangrijk onderdeel van ons leven geworden. Zo is 96% van de Nederlanders dagelijks actief

[2] Nu komt het heel af en toe toch nog wel eens voor dat mensen niet door hebben dat De Speld een satirisch nieuwsplatform is.

op sociale media (Wijker, 2020). Als gevolg daarvan consumeren we veel van onze informatie dan ook via deze sociale media platformen. Echter is er op sociale media weinig controle op wat voor informatie er wordt verspreid vanwege de aard van de platformen: iedereen genereert content en het platform verspreidt enkel de content. De bekende uitspraak "dat heb gestaan op Facebook" tijdens een uitzending van Eenvandaag (2015) geeft goed weer wat het resultaat kan zijn van mensen die feitelijk onjuiste informatie consumeren via sociale media platformen[3]. Hoe kan het toch dat *information disorders* een steeds groter fenomeen worden in ons digitale tijdperk? En, waarom lijken zo veel mensen zich slecht bewust te zijn van *information disorders*? In paragraaf 3.1 geef ik uitleg over het business model van sociale media platformen wat resulteert in een design dat gevoelig is voor *information disorders*. Vervolgens bespreek ik in paragraaf 3.2 de psychologische neigingen en biases waar de online bedrijven op inspelen. Ook bespreek ik hoe ons brein minder goed bestand is tegen *information disorders* dan we vaak initieel denken. Ten slotte bespreek ik in paragraaf 3.3 hoe algoritmes onze psychologische neigingen nog eens extra prikkelen.

3.1. De strijd om jouw aandacht

Wat veel gebruikers van sociale media platformen vergeten, is dat ze dagelijks gebruik maken van bedrijven. Het gebruik van sociale media platformen is nooit geheel gratis: er zit een heel businessmodel achter en jouw aanwezigheid op deze platformen is geld waard. En het grootste economisch goed wat te winnen valt? Jouw aandacht.

Veel onderzoekers hebben lang gedebatteerd over de precieze waarde van aandacht in nieuwe media, maar de meesten zijn het met elkaar eens; aandacht is de meest waardevolle grondstof in het moderne kapitalisme (Zulli, 2018: 139). Waar informatie overvloedig is, is aandacht schaars omdat de grenzen inherent zijn aan de neurofysiologie van de waarneming en er simpelweg beperkingen zijn aan de tijd die beschikbaar is (Terranova, 2012: 13). We spreken dan ook over een 'aandachtseconomie' omdat schaarste een belangrijk aspect is van een volwaardige economie (Ibid.).

Maar waarom is aandacht zo'n kostbaar goed voor sociale media platformen? Hoe langer de aandacht van mensen gevangen kan worden, hoe meer sociale media platformen verdienen. Professor Timothy Wu bedacht de term 'attention broker' om te beschrijven hoe

[3] Tijdens een uitzending van Eenvandaag (2015) over een protest tegen de komst van een AZC in Purmerend zegt een geïnterviewde protester dat ''haar pleegzoon van 21 geen werk kan krijgen terwijl de asielzoekers binnen 4 maanden werk krijgen''. Op de vraag hoe ze aan die informatie kwam, antwoordde de vrouw: ''dat heb gestaan op Facebook''. Deze uitspraak kreeg vervolgens veel kritische aandacht in de media.

bedrijven, zoals bijvoorbeeld Facebook en Twitter, de aandacht van consumenten trekken om vervolgens die aandacht door te verkopen aan adverteerders (Obear, 2018: 1007). Het businessmodel van dit soort online platformen is dan ook dat degene die de meeste aandacht weet te trekken en het langst die aandacht weet vast te houden ook de grootste financiële winst binnen haalt, omdat ze in staat zijn die aandacht door te verkopen als adverteerruimte. Daarnaast is het belangrijk voor adverteerders dat deze marketing relatief weinig kost maar het wel zeer effectief werkt (ze kunnen immers zo overstappen naar een ander platform). Ondertussen is Facebook hier grootmeester in geworden door zelfs de kleinste adverteerders de kans te geven om voor een relatief goedkope prijs advertenties te kunnen kopen, die vervolgens zeer doelgericht gepromoot kunnen worden op basis van de kenmerken van gebruikers. Facebook heeft een goed beeld van de 'kenmerken' van hun gebruikers door het gebruik van big data. Hierdoor heeft Facebook goed zicht op wat voor soort advertenties wel en niet effectief zijn en verzamelen ze steeds meer gebruikersinformatie over hoe advertenties doelgerichter ingezet kunnen worden (Obear, 2018: 1031). Zodoende verdienen sociale media platformen aan onze aandacht door een bemiddelende (attention broker) rol te spelen.

Het is dus van groot belang voor deze online bedrijven om steeds innovatievere manieren te bedenken om de aandacht van gebruikers te blijven vangen. De steeds vernieuwde mechanismen achter het design van sociale media platformen zorgen er dan ook voor dat iemands aandacht zo lang mogelijk gericht blijft op de content van het sociale media platform in kwestie. De mechanismen die sociale media platformen gebruiken om aandacht te trekken zijn onder andere cognitieve biases zoals verliesaversie (men wil liever een verlies vermijden dan winst ontvangen) en sociale vergelijking (denk hierbij aan 'likes') (Williams, 2018: 33). Zo stuurt bijvoorbeeld Instagram continue pushnotificaties die je oproepen dat iemand iets heeft gepost en je dat niet wil missen wanneer je een tijdje de applicatie niet hebt geopend (verliesaversie) en krijgt je continue meldingen wanneer je 'likes' hebt gekregen op een foto die je hebt gepost (sociale vergelijking).

Daarnaast is het gebruik van online platformen voor een groot deel gewoonte, een gegeven waar sociale media platformen maar al te goed van op de hoogte zijn. Consumenten van sociale media platformen starten met het gebruik van de service vanwege *beloningen*, of positieve uitkomsten, die ze behalen door gebruik te maken van het platform (Anderson & Wood, 2020: 85). Te denken valt aan 'likes' vergaren op een foto die je post of het kunnen chatten met een familielid die in een ander land woont. Als consumenten eenmaal gewoontes hebben ontwikkeld op het gebied van mediagebruik, komt de reactie (bijvoorbeeld de applicatie

Instagram openen) automatisch en onbewust in hen op wanneer zij bepaalde signalen waarnemen (bijvoorbeeld een bepaald geluid wanneer iemand je foto liket op Instagram) (Ibid.). Push notificaties zijn daarom een belangrijk mechanisme van sociale media platformen om onze aandacht te blijven lokken wanneer we niet op onze telefoon of tablet zitten.

Verder maken veel sociale media platformen gebruik van autoplay functies bij alle video's (nadat je video klaar is, springt er meteen een nieuwe video aan) en kun je eindeloos doorscrollen door je feed (Anderson & Wood, 2020: 89). Dit specifieke design speelt in op luiheid of uitstelgedrag en zorgt ervoor dat consumenten zo lang mogelijk blijven kijken en daarmee hun aandacht blijven focussen op de content van een sociale media platform. Daarbij is het voornamelijk cruciaal dat het design aanzet tot reageren en niet aanzet tot denken, omdat het aanzetten tot reageren doeltreffender is bij het bevorderen van gewoontevorming (Anderson & Wood, 2020: 90). Als het gaat om effectief de aandacht van consumenten te trekken en vast te blijven houden dan werken eindeloze filmpjes van katten die iets grappigs doen veel effectiever dan artikelen van de NRC. Zodra deze automatische reacties (push notificaties en eindeloze scrolmogelijkheden) er insluipen bij consumenten, verdwijnt de oorspronkelijke motivatie van consumenten in louter onbewuste gewoontes.

Er is echter nog iets wat een cruciale rol speelt in het design van sociale media platformen: het uitlokken van emotionele reacties. Een onderzoek van Brady et al. (2019) laat zien dat het gebruik van moreel- emotionele taal (woorden zoals schaamte, moord en haat) door invloedrijke mensen op sociale media sterk wordt geassocieerd met de toename van de verspreiding van die berichten. Wederom geldt hier: het design van sociale media platformen is zo gemaakt dat de content die beter en sneller verspreidt (in dit geval door het uitlokken van emoties) ook meer adverteerders oplevert. Het fenomeen 'clickbait' is hier een goed voorbeeld van. Clickbaits hebben als grootste doel om zo veel mogelijk aandacht te genereren middels titels die veel emotionele reacties uitlokken. De crux is echter dat clickbaits vaak gebruik maken van misleidende en soms zelfs feitelijk onjuiste informatie, aangezien dit resulteert in meer klikgedrag van mensen. Deze ontwikkeling is dus zeer gevoelig voor *information disorders*. Dat dit in de praktijk ook zo werkt, werd aangetoond in een tien jaar durend onderzoek van MIT. Het onderzoek had als conclusie dat feitelijk onjuiste informatie veel meer mensen kon bereiken dan feitelijk juiste informatie (Vosoughi et al., 2018: 3). Het grotere bereik van feitelijk onjuiste informatie wordt veroorzaakt doordat mensen *zelf* sneller geneigd waren om de feitelijk onjuiste informatie vaker te delen dan de feitelijk juiste informatie. De reden voor het grotere bereik had te maken met het emotionele taalgebruik en de 'nieuwigheid' (een

verrassingseffect) die meer verbonden blijkt te zijn met feitelijk onjuiste informatie dan feitelijk juiste informatie (Vosoughi et al., 2018: 5). Deze studie laat zien dat veel feitelijk onjuiste informatie bepaalde kenmerken heeft die er voor zorgen dat ze eerder de interesse van mensen kunnen prikkelen. Het is dan ook logisch dat er gretig gebruik van wordt gemaakt van deze kenmerken door personen die willen dat hun informatie de aandacht opvangt van zo veel mogelijk mensen. In de volgende paragraaf wordt er verder ingegaan op een verklaring van deze conclusie.

Het is nu duidelijk dat het business model van sociale media platformen is gebaseerd op het doorverkopen van jouw aandacht aan adverteerders. Om die reden is het noodzakelijk dat ze hun design zo inrichten dat consumenten geprikkeld blijven om hun aandacht zo lang mogelijk te blijven focussen op de content die wordt aangeboden. Doordat *information disorders* meer aandacht krijgen, levert het de platformen meer geld op waardoor deze online platformen een goede broedplaats zijn voor nepnieuws en misleidende informatie. De bovengenoemde processen achter sociale media platformen zijn cruciaal voor het begrijpen waarom bepaalde keuzes worden gemaakt over het design van de platformen vanuit een economisch oogpunt: het is namelijk rendabel om hun algoritmes en design zodanig in te stellen dat 'klanten' vooral blijven hangen, ook al gaat dat gepaard met het verspreiden van *information disorders*. In die zin is het verspreiden van *information disorders* een neveneffect van het business model en design van deze platformen die inspelen op onze psychologische biases. In de volgende paragraaf zoomen we wat meer in op de psychologie van mensen en waarom het design van sociale media platformen zo goed werkt.

3.2. Ons brein en information disorders

Het design van sociale media platformen zorgt ervoor dat *information disorders* makkelijk kunnen verspreiden. In deze paragraaf bespreken we hoe we aan de hand van theorieën uit de psychologie kunnen verklaren waarom we geneigd zijn om overtuigd te raken van foute en misleidende informatie.

Lang was het rationalistische model de focus van de psychologie en de filosofie als het ging om gedrag en ons oordelen verklaren (Haidt, 2000). We geloven namelijk maar al te graag dat we vooral rationele wezens zijn. De inzichten omtrent Dual Process-theorieën helpen ons echter veel beter verklaren waarom fenomenen als *information disorders* zo makkelijk op het internet floreren. Dual Process-theorieën onderzoeken hoe er twee soorten processen aan het werk zijn wanneer mensen denken en handelen. Het gaat hierbij om twee parallel lopende

systemen die tot verschillende conclusies kunnen komen (Haidt, 2001; Kahneman, 2011). In zijn wereldbekende boek 'Thinking Fast en Slow' legt psycholoog Daniel Kahneman op een vereenvoudigde manier uit hoe onze menselijke geest gebruik maakt van systemen die hij systeem 1 en systeem 2 noemt. Systeem 1 werkt automatisch en snel, terwijl systeem 2 aandachtig en met volle inspanning opereert (Kahneman, 2011: 20-21).

Het voordeel van systeem 1 is dat het weinig moeite kost doordat cognitieve heuristieken (regels die je brein toepast) ervoor zorgen dat we snel en efficiënt keuzes kunnen maken. Tegelijkertijd is het nadeel dan weer dat systeem 1 snel fouten maakt omdat er simpelweg weinig tot geen reflectie bij komt kijken (Ibid.). Systeem 2 daarentegen maakt minder fouten maar kost zo veel moeite dat het slechts gelimiteerd ingezet kan worden (Kahneman, 2011: 23). De samenwerking tussen de twee systemen is erg efficiënt: we gebruiken voornamelijk systeem 1 en op de momenten dat het lastiger wordt, springt systeem 2 bij (Kahneman, 2011: 25). De meeste informatie die wij dagelijks consumeren, verwerken we dus in eerste instantie met behulp van systeem 1. Iedere keer wanneer we nieuwe informatie ontvangen, zoekt systeem 1 automatisch in ons langetermijngeheugen naar bevestigend bewijs en genereert vervolgens een reactie in minder dan één seconde (Moravic et al., 2018: 8). Systeem 1 produceert daarmee een 'Feeling of Rightness' (een metacognitieve ervaring die voorspelt of intuïtieve reacties zullen veranderen) die bepaalt of we vertrouwen hebben in de conclusies (Ibid.). Systeem 1 is namelijk zo gemaakt dat het snel conclusies kan trekken (Kahneman, 2011: 209) en het corrigeren van deze intuïtieve voorspellingen is meer een opdracht voor systeem 2 (Ibid.: 192). Echter zoals we al eerder lazen, gebruiken we systeem 2 slechts gelimiteerd.

Het volgende experiment van Harvard (2020) toont de wisselwerking tussen systeem 1 en 2 goed aan. Zo bleek uit het experiment dat wanneer deelnemers gevraagd werd om actief te pauzeren en na te denken over of een titel juist of onjuist was, ze minder snel geneigd waren om de onjuiste informatie te verspreiden dan wanneer ze niet gevraagd werden om te pauzeren (Fazio, 2020). Wanneer we informatie consumeren wat onder *information disorders* valt, dan kost het relatief veel moeite om sceptisch te zijn over de informatie in kwestie. Of beter gezegd, het kost ons voornamelijk meer tijd en cognitieve aandacht – tijd en aandacht die we vaak niet willen of kunnen nemen als we op de automatische piloot van systeem 1 varen. Om die reden is de hoeveelheid *information disorders* wat zich verspreid op sociale media platformen zorgelijk: ons brein is er simpelweg niet op gemaakt om altijd sceptisch en rationeel te zijn als het gaat om informatie verwerken.

Bovendien speelt er ook iets anders mee wat bepalend is voor de mate waarin we informatie accepteren als waarheid: onze intuïties. Wanneer we informatie lezen die valt onder *information disorders* dan is de kans dat we het geloven groter wanneer het onze intuïties (emotionele reacties) rechtvaardigt (Haidt, 2001). Psycholoog Jonathan Haidt verwerpt het idee dat we onze oordelen baseren op redenaties en reflectie. Op basis van empirische onderzoeken beargumenteert hij dat we de meeste oordelen maken op basis van intuïties en dat we vaak pas gebruik maken van redeneren en reflecteren wanneer we onze intuïties willen verdedigen, of te wel post hoc redeneren. Volgens Haidt moeten we ons oordelen daarom meer zien als een advocaat die een cliënt verdedigt dan als een rechter die de waarheid zoekt (Haidt, 2001: 10). Deze metafoor is makkelijk te vertalen naar hoe mensen sneller geneigd zijn om online informatie te zoeken die aansluit op bestaande geloofsovertuigingen (de advocaat) in plaats van dat ze verschillende bronnen zullen vergelijken om tot een goede eindconclusie te komen (de rechter). Ook Kahneman (2011) schrijft dat wanneer het om een houding gaat, systeem 2 meer de 'apologeet' is dan de 'criticus' ten aanzien van (de emoties die voortkomen uit) systeem 1 (103). Wanneer we informatie consumeren die onze intuïties rechtvaardigen, zijn we dus sneller geneigd om het te accepteren. Vervolgens zijn we pas geneigd om onze redenatie en reflectie in te zetten voor het verdedigen van onze intuïties.

Dit proces wordt bovendien extra versterkt door de zogenaamde 'confirmation bias': men consumeert graag informatie die zijn of haar geloofsovertuigingen en verwachtingen bevestigen (Nickerson, 1998). Wanneer een persoon die gelooft dat er is gefraudeerd tijdens de verkiezingen van 2020 een video op Facebook ziet waarbij er iets verdachts lijkt te gebeuren met stembiljetten, dan is de kans groter dat deze persoon het filmpje geloofwaardig vindt. Wanneer vervolgens een persoon die *niet* gelooft dat er is gefraudeerd tijdens de verkiezingen dezelfde video ziet, zal deze persoon veel sneller een sceptische houding aannemen tegenover de betrouwbaarheid van de video: is deze video wel recent gemaakt en is de video wel gemaakt in de VS? Informatie over de schadelijkheid van vaccinaties zal sneller opgemerkt en verspreid worden door anti-vaxxers dan door mensen die geloven in de werking en noodzaak van vaccines. Een onderzoek van Moravic et al. (2018) laat zien dat zelfs wanneer er waarschuwingslabels bij berichten staan, men het bericht blijft geloven wanneer de inhoud valt binnen hun a priori geloofsovertuigingen (20-21). Bovendien speelt 'wishful thinking' ook een rol: wanneer informatie inhoud bevat waardoor iemand er beter vanaf komt, dan zijn mensen sneller geneigd het te geloven (Mayraz, 2011). Kortom, *information disorders* worden in stand gehouden omdat we ze vaak graag *willen* geloven en – zoals we al eerder bespraken – omdat

we voornamelijk leunen op de quasi-automatische processen van systeem 1, die automatisch in ons langetermijngeheugen naar bevestigend bewijs zoeken. Bij deze processen kan het dus voorkomen dat er sprake is van biases, zoals de 'confirmation bias' of 'wishful thinking'. Het kan echter ook zo zijn dat je niet per se op clickbait klikt omdat je het graag wil geloven, maar ook simpelweg omdat het gebruik maakt van emotionele taal en/of een verassingseffect (zoals we besproken hebben in 3.1).

Concluderend kan er gesteld worden dat onze psychologische neigingen een perfecte match zijn met het business model en design van sociale media platformen (en dat weten de bedrijven achter de online platformen ook). Het resultaat van deze goede match is dat veel mensen makkelijk beïnvloed raken door de content – die niet noodzakelijk met waarheid te maken heeft – die verschijnt op de diverse platformen. Ten eerste omdat we informatie verwerken aan de hand van systeem 1 waardoor we vaak niet erg kritisch zijn op de informatie die we tot ons nemen (en dus gevoelig zijn voor emotioneel taalgebruik bijvoorbeeld) en ten tweede omdat we informatie verwerken aan de hand van onze intuïties en heuristieken (wat samenhangt met systeem 1) waardoor we sneller geneigd zijn informatie overtuigender te vinden wanneer het aansluit op bestaande geloofsovertuigingen, of te wel onze confirmation bias.

3.3. Hoe filterbubbels er nog een schepje bovenop doen

In de vorige twee paragrafen heb ik uitgelegd waarom het business model - en daarmee het design - van sociale media platformen goed inspelen op onze psychologische neigingen. In deze paragraaf gaan we nog een stapje verder: algoritmes van sociale media platformen versterken namelijk de psychologische neigingen die besproken zijn in paragraaf 3.2.

In 2011 introduceert Eli Pariser in zijn boek 'The Filter Bubble: What the Internet Is Hiding from You' het concept 'filterbubbel'. Pariser betoogt in zijn boek dat online platformen hun aangeboden informatie per gebruiker personaliseren. De gepersonaliseerde informatie wordt gebaseerd op eerdere zoekgeschiedenis, iemands voorkeuren en iemands contacten binnen het sociale media platform (Pariser, 2011). Pariser's 'filterbubbel' houdt in dat we door deze gefilterde informatie in een bubbel terecht komen waarbij we te maken krijgen met onzichtbare 'autopropaganda' dat ons indoctrineert met onze eigen ideeën (Pariser, 2011: 13). Het is overigens niet nieuw dat we slechts media consumeren wat past bij onze interesses en geloofsovertuigingen: we selecteren en negeren altijd al bepaalde informatie omdat we simpelweg nooit alle informatie kunnen verwerken. Wat de filterbubbel echter ongewoon

maakt, is dat het ons introduceert aan drie nieuwe ontwikkelingen waar we nog niet eerder mee te maken hebben gehad (Pariser, 2011: 10).

Ten eerste, jij bent de enige persoon in de filterbubbel. Algoritmes zijn zo gepersonaliseerd dat niemand exact dezelfde data voorgeschoteld krijgt bij een zoektocht op het internet. Ten tweede, jij hebt zelf niet door dat je in een filterbubbel zit omdat het onzichtbaar is. Algoritmes zijn namelijk gepersonaliseerd maar je weet zelf niet wat de criteria voor de personalisatie zijn (laat staan wat de personalisatiecriteria voor anderen zijn) waardoor je geen idee hebt hoe jouw filterbubbel er uit ziet. Tegelijkertijd werkt het ook als een soort vicieuze cirkel: door bepaalde informatie weer te geven of te blokkeren, heeft de filterbubbel invloed op de keuzes die je (online) maakt en helpt het je persoonlijkheid vorm te geven (Pariser, 2011: 64). Ten derde, je kiest er niet zelf voor om wel of niet in de filterbubbel te stappen. Wanneer je de Volkskrant of de Telegraaf koopt, dan kies je actief voor het soort nieuws dat je wil consumeren. Dit proces gebeurt daarentegen niet wanneer je nieuws aangeboden krijgt op Facebook: niet jij zelf, maar algoritmes bepalen wanneer en wat je te zien krijgt op het internet. Kortom, het is nu eenmaal onvermijdelijk om niet een selectie te maken van alle informatie die je tot je kunt nemen, maar in het geval van de filterbubbel wordt de keuze van die selectie volledig op maat gemaakt zonder dat je daar invloed op hebt.

In eerste instantie lijkt de filterbubbel wellicht een vrij onschuldige ontwikkeling: het kan immers best handig zijn dat je enkel informatie te zien krijgt waar je waarschijnlijk toch al naar op zoek was. Maar wat goed werkt voor consumenten, hoeft echter niet goed te werken voor burgers in een democratie (Pariser, 2011: 15). Volgens Pariser kleeft er namelijk een grote sociale consequentie vast aan de manier waarop we ons internet vormgeven: onze democratie is in gevaar. Democratie vereist immers twee dingen: burgers zijn in staat elkaars standpunten in te zien en burgers kunnen aan dezelfde bronnen en feiten komen (Pariser, 2011: 4). Beide vereisten zijn moeilijk haalbaar in het huidige internet landschap. Zo zien we door onze filterbubbels juist minder vaak radicaal andere standpunten. Daarnaast zien we geen gedeelde informatie omdat de informatie grotendeels gepersonaliseerd is (Ibid.). Met de filterbubbels en de verspreiding van *information disorders* komen sociale media platformen dus ineens in het vaarwater van onze democratie. En we zien daar nu al de problemen van in: want als een democratie een publiek debat veronderstelt, dan wordt het publieke debat nu ondermijnt omdat (1) mensen in kleinere groepen met elkaar praten en (2) ook nog eens op basis van informatie die *information disorders* bevat. In paragraaf 4.2 ga ik hier uitgebreider op in.

De filterbubbel leidt tot veel discussies over of dit fenomeen überhaupt bestaat en of het daadwerkelijk zo gevaarlijk is als dat Pariser claimt. Critici benadrukken hoe effectieve personalisatie juist een belangrijke rol kan spelen in het verspreiden van nieuws, zoals bijvoorbeeld de rol van Twitter in de revoluties in het Midden-Oosten (Weisberg, 2011). Onderzoekers concludeerden dat er nog te weinig empirisch bewijs is die de bezorgdheid over filterbubbels rechtvaardigt (O'Hara & Stevens, 2015; Zuiderveen Borgesius et al., 2016). Tegelijkertijd concluderen andere onderzoekers dat selectieve blootstelling aan informatie wel degelijk aan het werk lijkt te zijn, wat leidt tot situaties waarin gebruikers van sociale media zich nauwelijks meer bezighouden met afwijkende informatie of standpunten (Grömping, 2014; Quattrociocchi et al., 2016). Bovendien tonen recentere experimenten aan dat er wel filterbubbels ontstaan en dat sociale media platformen vatbaar zijn voor extreme polarisatie (Chitra & Musco, 2020). Een sociale media platform waarbij het idee van filterbubbels meerdere keren duidelijk is aangetoond, is YouTube. Zo creëerde het algoritme van YouTube een steeds groter wordende extreemrechtse community (Bryant, 2020; Lewis, 2018; Tokmetzis et al., 2019). Onderzoek is dan niet eenduidig over de mate van invloed van filterbubbels door sociale media, maar we dansen als samenleving hoe dan ook op een koord als het gaat om internet en onze democratie. Enerzijds is het internet een zegen (we hebben nog nooit zo makkelijk zo veel informatie kunnen consumeren) voor onze democratie en anderzijds een vloek (we denken dat we alle informatie zien, maar dat is niet meer het geval door personalisatie als gevolg van algoritmes) voor onze democratie.

Het voorbeeld van YouTube toont goed aan hoe onze psychologische neigingen, besproken in paragraaf 3.2, extra versterkt worden door algoritmes. Zodra gepersonaliseerde informatie zich steeds meer aanpast op wat iemand graag *wil* zien, dan wordt het overtuigende effect ook steeds heftiger. Mensen belanden daarmee in een vicieuze cirkel: je krijgt de informatie voorgeschoteld die je wil zien en daarmee lijkt die informatie ook steeds overtuigender omdat je het simpelweg steeds vaker ziet, ook wel 'repetition bias' genoemd (Bodner et al., 2006). Onze confirmation bias speelt dus een grote rol, maar het algoritme doet er nog een extra schepje bovenop door als het ware mee te gaan in onze confirmation bias zonder dat we ons daar actief bewust van zijn.

Bovenstaande informatie is op zichzelf al problematisch maar het wordt vooral problematisch wanneer mensen terecht komen in filterbubbels waar voornamelijk *information disorders* worden verspreid. Mensen raken namelijk extreem overtuigd van feiten die onjuist zijn. Tot een zekere mate kan deze ontwikkeling vrij onschuldig zijn, maar het kan gevaarlijk

worden wanneer andere mensen geschaad worden. Denk hierbij aan ouders die weigeren hun kinderen te vaccineren door informatie uit artikelen op het internet. De groepsimmuniteit kan verdwijnen waardoor bepaalde ziektes terug kunnen komen. Kinderen die om medische redenen niet gevaccineerd mogen worden, raken in gevaar doordat de groepsimmuniteit hen niet meer kan beschermen. Een ander voorbeeld zijn de heftige geweldsgolven tegen de Rohingya-bevolking in Myanmar. Zo gebruikten machthebbers Facebook om feitelijk onjuiste berichten te verspreiden die aanzetten tot massageweld, verkrachting en zelfs moord (Stevenson, 2018).

Sociale media platformen spelen in op onze psychologische neigingen vanwege financieel gewin wat resulteert in potentiële broedplaatsen voor *information disorders*. Door het invoeren van algoritmes doen de online bedrijven er nog een schepje bovenop. Ter conclusie kan er gesteld worden dat sociale media platformen de verspreiding van *information disorders* faciliteren. Niet omdat ze de informatie zelf creëren, maar wel omdat hun business model en design zeer gevoelig zijn voor het verspreiden van *information disorders*. Daarom analyseren we vervolgens in welke mate deze sociale media platformen een morele verantwoordelijkheid dragen betreffende het verspreiden, screenen en/of censureren van *information disorders*? Want als *information disorders* potentieel mensen ernstig kunnen schaden, dan wordt het tijd dat we de morele verantwoordelijkheidsvraag gaan stellen.

4. De morele verantwoordelijkheid van sociale media platformen

In het vorige hoofdstuk heb ik uitgelegd waarom sociale media platformen een broeiplaats zijn voor *information disorders.* Dragen sociale media platformen een verantwoordelijkheid om te zorgen dat *information disorders* niet de boventoon voeren? Dit hoofdstuk pleit als antwoord: ja. Niet alleen zorgt de slechte controle op *information disorders* ervoor dat ze zich makkelijk kunnen verspreiden, sociale media platformen faciliteren zelfs de verspreiding van *information disorders*. In paragraaf 4.1 onderzoek ik of we sociale media mogen beschouwen als morele actoren. Deze stap is immers belangrijk om als eerste te maken wanneer we willen spreken over morele verantwoordelijkheid. Vervolgens behandel ik in paragraaf 4.2 op welk ethisch principe (de meerwaarde van waarheidsvinding) we sociale media platformen verantwoordelijk moeten houden.

4.1. Zijn sociale media platformen morele actoren?

Over de vraag of we personen, laat staan bedrijven en collectieven, verantwoordelijk kunnen houden voor hun handelen, kan een geheel nieuwe scriptie geschreven worden. Morele verantwoordelijkheid is een complex onderwerp waar geen consensus over is. Het debat met betrekking tot vrije wil en de diverse standpunten binnen die discussie impliceren verschillende uitgangsposities voor het debat omtrent morele verantwoordelijkheid. Daarom wil ik graag duidelijk maken dat deze scriptie niet verder ingaat op het vrijewilsdebat. Ik onderzoek in deze paragraaf van naderbij hoe we aan kunnen tonen dat sociale media platformen verantwoordelijk gehouden kunnen worden en waar die verantwoordelijkheid precies in kan bestaan.

Dat we bedrijven als handelingsbekwaam beschouwen, zien we terug in het feit dat we bedrijven als rechtspersonen behandelen. Wanneer iets als handelingsbekwaam wordt geacht, dan wordt er ook een zekere verantwoordelijkheid verwacht. Sinds de tweede helft van de 20e eeuw wordt de term 'Maatschappelijk verantwoord ondernemen' (MVO) steeds vaker gebruikt om aan te geven dat bedrijven een bepaalde verantwoordelijkheid hebben ten opzichte van de samenleving (Garriga & Melé, 2004). Keith Davis was een van de eersten die onderzocht wat de macht en impact van bedrijven is op de samenleving (Garriga & Melé, 2004: 55; Davis, 1960). Hij nam 'sociale macht' dan ook als uitgangspunt om te beargumenteren dat bedrijven verantwoordelijkheid dragen voor de samenleving: ''responsibility goes with power'' (Davis, 1960: 71). Met andere woorden, de verantwoordelijkheid van bedrijven is inherent aan de macht en impact die ze hebben op anderen. Dit sluit goed aan bij de situatie die geschetst wordt in deze scriptie[4]. Aangezien sociale media platformen een leidende rol spelen bij de verspreiding van *information disorders*, moeten hun bestuurders zich realiseren dat een weloverwogen en innovatief beleid noodzakelijk is om de maatschappelijke vraagstukken die door de ontwikkeling van hun platform worden veroorzaakt aan te pakken. En minstens even belangrijk: als liberale samenleving moeten we die morele verantwoordelijkheid van bedrijven ook eisen als ze zo veel invloed op ons dagelijks leven en onze democratie hebben.

Om te bepalen of een bedrijf impact heeft op de samenleving en daarmee ook een morele verantwoordelijkheid met zich meedraagt, kunnen we ons het beste baseren op bepaalde voorwaarden. Van de Poel et al. (2015) formuleren vijf concrete voorwaarden om van een

[4] Er is veel literatuur te vinden over MVO. De verschillende theorieën zijn grofweg in te delen op vier verschillende focuspunten: economie, politiek, sociale integratie en ethiek (Mélé, 2008). Op basis daarvan zijn de vier grootste theorieën de 'Shareholder Value Theory', de 'Corporate Citizenship Theory' , de 'Corporate Social Performance' en de 'Stakeholder Theory'. Deze scriptie maakt echter gebruik van deze theorieën omdat de focus niet ligt op bedrijfsethiek maar meer op het algemeen kijken naar de voorwaarden van moreel actorschap.

morele actor te kunnen spreken: capaciteit, causaliteit, kennis, vrijheid en kwaad-doen (21). Ik kies voor deze vijf voorwaarden omdat het een relatief simpele en algemene manier is om moreel actorschap te toetsen bij individuen maar ook collectieven. Aan de hand van deze voorwaarden gaan we daarom analyseren of we sociale media platformen als morele actor kunnen beschouwen en daarnaast, meer specifiek, of we ze moreel verantwoordelijk kunnen houden omtrent *information disorders*.

De eerste voorwaarde is *capaciteit*. Deze voorwaarde is meteen de meest controversiële voorwaarde als het gaat om collectieven: kun je stellen dat een groep mensen gezamenlijk de capaciteit bezit om verantwoordelijk te handelen en daardoor als groep aansprakelijk is? Wel wanneer je als extra voorwaarde neemt dat de capaciteit gerelateerd is aan de besluitvorming van het collectief (González, 2002: 103; Van de Poel et al., 2015: 57). Hierbij is het belangrijk dat (1) er capaciteit is om morele redenatie te gebruiken in de besluitvorming van het collectief en (2) dat het besluitvormingsproces niet alleen de zichtbare handelingen beheert maar ook de interne structuur van beleid en regels (González, 2002: 103). De bedrijven achter sociale media platformen zijn wel degelijk in staat (ergo capaciteit) om in te grijpen bij content dat *information disorders* bevat omdat zij uiteindelijk de macht hebben om te monitoren wat wel en niet wordt verspreid. Bovendien zijn ze als collectief -bestuurders, medewerkers en stakeholders- in staat om moreel te reflecteren op de consequenties van bepaalde beleidskeuzes.

De volgende voorwaarde, *Causaliteit,* houdt in dat de morele actor een situatie heeft veroorzaakt en het dus terug te traceren is of diegene aansprakelijk is (Van de Poel et al., 2015: 22). Hierbij gaat het niet per definitie om een slechte intentie: het gaat simpelweg om het causale verband tussen een handeling en een uitkomst waarbij anderen schade wordt toegebracht. In de paragrafen 3.1, 3.2 en 3.3 is er geconcludeerd dat sociale media platformen een faciliterende rol spelen betreffende het verspreiden van *information disorders* omdat (1) *information disorders* aantrekkelijk zijn binnen hun businessmodel en (2) hun design *information disorders* in de hand werkt. Hierbij is er geen sprake van een evidente intentie voor het *willen* verspreiden van *information disorders*, maar er is wel sprake van een causaal verband tussen sociale media platformen en de negatieve consequenties (de effecten van *information disorders*) voor de samenleving. In andere woorden, het is door het bewuste design van de platformen (dat gebaseerd is op hun kennis van de menselijke brein) dat verspreiding van *information disorders* (met alle gevolgen van dien) gefaciliteerd wordt. Zodoende kan er gesteld worden dat deze bedrijven (mede) verantwoordelijk zijn voor de negatieve implicaties. Een ander design zou namelijk andere gevolgen hebben. Zo zou het veranderen van de algoritmes er bijvoorbeeld al

voor kunnen zorgen dat mensen niet in een vicieuze cirkel terecht komen met hun confirmation bias.

Vervolgens betekent de voorwaarde *kennis* dat een actor een handeling vrijwillig heeft uitgevoerd zonder dwang of onwetendheid (Ibid.). Er kan wel sprake zijn van onwetendheid over de negatieve consequenties van het businessmodel van sociale media platformen in de vroegere stadia, maar na de vele onderzoeken en evidente voorbeelden valt er niet meer te verschuilen achter onwetendheid. Anders gezegd kan er ook altijd met terugwerkende kracht ingegrepen worden nadat er wel kennis wordt genomen van de consequenties. Zo heeft YouTube, ter illustratie, aanpassingen gemaakt in hun algoritme nadat er uit meerdere onderzoeken bleek dat mensen sneller radicaliseerden door het toenmalige algoritme dat bepaalde welke video's werden voorgeschoteld aan mensen (YouTube, 2019). Kortom, sociale media platformen hebben kennis over de effecten van hun handeling, al is het met terugwerkende kracht.

De voorwaarde *vrijheid* staat uiteraard voor het vrij kunnen handelen als morele actor (Van de Poel et al., 2015: 22-23). Vrijheid kan echter op verschillende wijze worden geïnterpreteerd. Vrijheid betekent in ieder geval dat er geen sprake is geweest van dwang. Daarnaast valt onder deze voorwaarde het debat over vrije wil, of te wel: hebben wij een vrije wil om keuzes te maken of zijn we gedetermineerd? Maar zoals reeds vermeld, zal deze scriptie niet verder in gaan op het vrijewilsdebat en wordt er uitgegaan van een vrijheid om moreel te kunnen handelen. De keuze voor een bepaald businessmodel binnen een bedrijf is een vrije keuze. Ook is er sprake van een vrije keuze om wel of niet in te grijpen als bedrijf nadat blijkt dat bepaalde keuzes of beleid negatieve consequenties meebrengen. Sociale media platformen voldoen om die reden aan de voorwaarde vrij te kunnen handelen.

De laatste voorwaarde is *kwaad-doen*. Wanneer we actoren als verantwoordelijk beschouwen, dan is er iemand schade aangedaan of is er een norm overtreden (Ibid.). Uiteraard heeft iemands ethische overtuiging invloed op wat we als 'kwaad-doen' beschouwen. Wat echter vooral van belang is in deze context, is dat er voldoende overeenstemming bestaat over het feit dat een actor verantwoordelijk wordt gehouden omdat de actor geacht wordt iets gedaan te hebben wat onwenselijke consequenties tot gevolg heeft. Deze voorwaarde is vrij makkelijk toe te passen in de casus van sociale media platformen. Er is sprake van de voorwaarde 'kwaad-doen' omdat er negatieve consequenties (schade aan individuen en democratie) zijn aan te duiden die terug te leiden zijn naar handelingen van sociale media platformen. In paragraaf 3.3

werden er twee concrete voorbeelden (de antivaxxers en de geweldsgolven in Myanmar) gegeven waaruit blijkt dat er daadwerkelijk sprake is van schade.

Aan de hand van deze vijf voorwaarden kan er geconcludeerd worden of een individu of collectief moreel verantwoordelijk gehouden kan worden. Door de voorwaarden capaciteit, causaliteit, kennis, vrijheid en kwaad-doen toe te passen op de bedrijven die sociale media platformen designen en hun rol in het faciliteren van information disorders, kan geconcludeerd worden dat ze morele actoren zijn en moreel verantwoordelijk gehouden kunnen worden voor de eventuele schade die hun platformen veroorzaken wanneer ze het verspreiden van *information disorders* in de hand werken.

4.2. De meerwaarde van waarheidsvinding in een democratie

Naast dat we hebben vastgesteld dat sociale media platformen beschouwd kunnen worden als moreel actor, is het ook van belang dat we een ethisch principe definiëren waarvoor we sociale media platformen verantwoordelijk willen houden. Mijn analyse van morele verantwoordelijkheid is tot dusver nog formeel gebleven: kan iemand of iets verantwoordelijk gehouden worden? Echter moet er dan ook vastgesteld worden wat die verantwoordelijkheid dan precies inhoudt. Want wanneer kunnen we stellen dat iemand moreel problematisch bezig is? We hebben daarom een criterium of ethisch principe nodig waaruit we moreel problematisch handelen kunnen onderscheiden van moreel onproblematisch handelen. Gezien de focus van deze scriptie, namelijk *information disorders* en hun effecten op mensen en de samenleving als geheel, is het een logische stap om het volgende principe te formuleren: de meerwaarde van waarheidsvinding in een democratie en publiek debat. Deze paragraaf onderzoekt waarom waarheidsvinding meerwaarde heeft en waarom het een probleem is wanneer sociale media platformen waarheidsvinding ondermijnen.

Volgens John Stuart Mill is het idee van waarheidsvinding cruciaal voor een goedwerkende samenleving. Mill beargumenteert dat wanneer we zo veel mogelijk opinies naast elkaar kunnen leggen, we het dichtst bij de waarheid kunnen komen aangezien opinies vaak maar een gedeelte van de waarheid met zich meebrengen (Mill, 2009: 77). Het streven naar waarheidsvinding zou vervolgens de gehele mensheid vooruithelpen op meerdere culturele en technologische aspecten. Mill vertrouwt op een proces van een redelijk en fatsoenlijk debat waarbij er een interactieve uitwisseling is tussen ideeën en argumenten. De relevante inzichten van Mill over waarheidsvinding (en vrijheid van meningsuiting) bespreken we uitgebreider in hoofdstuk 5.

Ook Jürgen Habermas pleit voor het bij elkaar brengen van opinies binnen een 'publieke sfeer'. Met het concept 'publieke sfeer' doelt Habermas op een ruimte waarbinnen een publieke opinie gevormd kan worden zonder dwang en het liefst door middel van een rationele discussie die open is voor alle burgers (Habermas et al., 1964: 49). Volgens Habermas is de publieke sfeer, die opereert als intermediair tussen staat en samenleving, één van de drie elementen (naast de autonomie van burgers en de inclusie van vrije en gelijke burgers) van een moderne democratie (Habermas, 2006: 412). Habermas richt zich op empirische onderzoeken die aantonen dat wanneer je groepen mensen met elkaar in gesprek laat gaan over bepaalde (controversiële) onderwerpen, het proces van het overleg niet in polarisatie maar zich juist meer in de richting van consensus ontwikkelt (Habermas, 2006: 414).

Habermas' idee sluit aan bij de theorie van deliberatieve democratie. Deze theorie verwijst namelijk naar een specifieke vorm van participatie binnen een democratie: discussies tussen geïnformeerde individuen over zaken die hen aangaan, leiden uiteindelijk tot een zekere vorm van consensus en collectieve besluiten (Wright & Street, 2007: 850). Een deliberatieve democratie heeft een potentieel 'waarheidsvindinggehalte' (Habermas noemt dit de epistemische dimensie van deliberatie) in de zin dat het de voorwaarden biedt om dichter bij de waarheid te komen (Habermas, 2006: 413-415). Waarheid moet in deze context niet worden opgevat als iets absoluut of objectief, maar als iets intersubjectief. Wat cruciaal is bij waarheidsvinding is het proces, dus de manier waarop discussies mogelijk gemaakt en vorm gegeven worden opdat we overeenstemming kunnen vinden en zo dichter bij de waarheid geraken. Habermas lijkt namelijk een onderscheid te maken tussen waarheid als feitenkennis en de zoektocht naar gedeelde waarden om onze samenleving op te baseren. In het geval van Mill, Habermas en de theorie van deliberatieve democratie is de beste vorm van waarheidsvinding het faciliteren van discussies waarbij mensen elkaar kunnen overtuigen van hun opinie met als uiteindelijke doel meer redelijkheid en overeenstemming. Het overtuigen gaat dan aan de hand van argumenten die goed onderbouwen waarom het juist is. Dus niet: valse zaken beweren door te charmeren met clickbaits.

Maar wat als de discussie in de publieke sfeer wordt verstoord of zelfs wordt gedomineerd door *information disorders*? Mill en Habermas vertrouwen op het proces van een redelijk en fatsoenlijk debat, maar dat debat kan steeds lastiger plaatsvinden op sociale media platformen: mensen kunnen minder open en rationeel argumenteren wanneer ze vastzitten in een filterbubbel en continu informatie te zien krijgen die ze misschien wel leuk vinden maar die anderen niet te zien krijgen en die vaak ook nog feitelijk onjuist is. Als liberale,

democratische samenleving moet ons dit zorgen baren. Voor de goede werking van een democratie, is er namelijk een publiek debat nodig, waarbij relevante informatie gedeeld wordt (1) door verschillende partijen en (2) die betrouwbaar is. Naast dat goed geïnformeerde mensen noodzakelijk zijn voor een goedwerkende democratie zorgen *information disorders* ook voor chaos in ons vertrouwen: wie kunnen we wel en niet vertrouwen? Kortom, het tast ons vermogen aan om betrouwbare informatie te delen met elkaar en op die manier gezamenlijk een discussie te voeren over beleid en de richting die we als samenleving op willen.

In de vorige paragraaf is er beargumenteerd waarom de bedrijven achter de sociale media platformen als moreel actor beschouwd kunnen worden en waarom we dat als samenleving ook zouden moeten doen: ze hebben veel impact op onze samenleving. Sociale media platformen moeten hun morele verantwoordelijkheid gaan dragen voor waarheidsvinding, dat wil zeggen dat ze moeten voorkomen dat/ingrijpen als *information disorders* de boventoon voeren aangezien *information disorders* momenteel juist waarheidsvinding ondermijnen. Een oplossing die sociale media platformen tot nu toe gebruiken om *information disorders* te bestrijden is het censureren van bepaalde content. Echter creëert die oplossing een nieuw vraagstuk dat behandeld dient te worden: in hoeverre is het censureren van *information disorders* een aanvaardbare schending van vrijheid van meningsuiting? En is er hier sprake van een spanning tussen enerzijds het principe vrijheid van meningsuiting en anderzijds het principe waarheidsvinding?

5. Vrijheid van meningsuiting en waarheidsvinding

De verspreiding van *information disorders* heeft epistemisch schadelijke effecten voor individuen en voor de gehele samenleving als democratie. Het design van sociale media platformen faciliteert de verspreiding van *information disorders* waardoor het principe ‘waarheidsvinding’ op dit moment wordt ondermijnd. Gegeven deze schadelijke effecten en het ondermijnen van waarheidsvinding hebben sociale media platformen een morele verantwoordelijkheid om in te grijpen bij content dat valt onder *information disorders*. Een oplossing tegen *information disorders* die sociale media platformen gebruiken, is het censureren van de desbetreffende content. In dit laatste hoofdstuk wordt aan de hand van Mill onderzocht of het censureren van *information disorders* een (on)aanvaardbare schending is van

vrijheid van meningsuiting en of er daadwerkelijk een spanning is tussen het principe ‘vrijheid van meningsuiting’ en ‘waarheidsvinding’. Ten slotte bespreken we oplossingen die ervoor zouden moeten zorgen dat waarheidsvinding niet wordt ondermijnd door sociale media platformen.

5.1. Vrijheid van meningsuiting en waarheidsvinding volgens Mill

Volgens artikel 19 van de Universele Verklaring van de Rechten van de Mens heeft ieder mens het recht om zich vrij uit te kunnen spreken en dus informatie te versturen en te ontvangen: ‘’Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart *information* and ideas through any media and regardless of frontiers’’ (United Nations, 2021). De term ‘vrijheid van meningsuiting’ heeft echter al een lange geschiedenis achter de rug voordat het werd erkend als een fundamenteel mensenrecht. De vader van het liberalisme, filosoof John Stuart Mill, wordt vaak het meest geassocieerd met vrijheid van meningsuiting. In deze sectie zoek ik dan ook betekenis en antwoorden in zijn tekst ‘On Liberty’ (2009). In paragraaf 4.2 is Mill al kort geïntroduceerd, maar in deze paragraaf gaan we wat dieper in op de relevantie van zijn tekst.

Het doel van Mills essay is uitzoeken waar de grens van macht ligt die de samenleving kan hebben over een individu (Mill, 2009: 4). Voor Mill is het vrij na kunnen denken onlosmakelijk verbonden met vrijheid van spreken en schrijven (Mill, 2009: 26). Zodoende schrijft hij dat een samenleving pas echt vrij is, ongeacht de vorm die de overheid aanneemt, wanneer men vrij is in zichzelf uiten (Mill, 2009: 23). Belangrijk is echter om te begrijpen dat Mill vrijheid van meningsuiting niet per se beargumenteert vanuit een individueel *recht* dat men zou hebben. Integendeel, hij neemt in zijn tekst als eerste beginsel utiliteit; hoeveel brengt het ons als samenleving uiteindelijk op? Zo schrijft hij het volgende: ‘’But the peculiar evil of silencing the expression of an opinion is, that is robbing the human race’’ (Mill, 2009: 29). Hieruit blijkt dat Mill zijn argument voor vrijheid van meningsuiting baseert op maatschappelijk nut: we gaan als gehele mensheid vooruit wanneer iedereen een steentje mag en kan bijdragen door zijn opinies te uiten. Het is cruciaal om nogmaals te benadrukken dat een opinie juiste informatie nodig heeft willen we goed geïnformeerd zijn als samenleving.

Het argument dat Mill maakt voor vrijheid van meningsuiting gaat over het streven naar waarheid. Alleen door zo veel mogelijk opinies naast elkaar te leggen kunnen we het dichtst bij de waarheid komen. Mill beredeneert namelijk dat opinies vaak maar een gedeelte van de waarheid met zich meebrengen en vrijwel nooit de gehele waarheid (Mill, 2009: 77). Daarnaast, stelt Mill, zijn wij als mensen behoorlijk feilbaar (Mill, 2009: 42). We maken nu

eenmaal vaak fouten, zo blijkt uit talloze voorbeelden uit de geschiedenis. Om vervolgens een sterke basis te kunnen bouwen voor waarheid, moeten we de desbetreffende basis altijd goed evalueren; de waarheid gedijt namelijk alleen wanneer het uitvoerig besproken wordt, anders blijft het louter een 'dood dogma' (Mill, 2009: 58). Met andere woorden: we zijn samen, na deliberatie, slimmer dan alleen, zonder deliberatie. Dat betekent echter niet dat er geen plaats is voor foute opinies, want: foute opinies zullen uiteindelijk worden bijgesteld doordat de juiste opinies verdedigd zullen worden (Mill, 2009: 29). Mill vertrouwt dus op het proces van deliberatie, net zoals Habermas (die we bespraken in 4.2).

Zit er dan geen beperking aan vrijheid van meningsuiting? Volgens Mill zit er wel degelijk een limitatie aan onze vrijheid van meningsuiting. Zeer expliciet schrijft hij in zijn tekst dat men nooit schade mag toebrengen aan anderen: ''The liberty of the individual must be thus far limited; he must not make himself a nuisance to other people'' (Mill, 2009: 94). Wanneer iemand dus wel schade toebrengt aan anderen dan mag er wel ingegrepen worden (Mill, 2009: 18). Dit principe is later bekend geworden als het 'schadebeginsel'.

Kortom, Mill pleit in zijn beroemde boek 'On Liberty' voor vrijheid van meningsuiting. Hetgeen brengt, volgens hem, de mensheid de meeste utiliteit: door iedereen de kans te geven zich te uiten, komen we steeds een beetje dichter bij de waarheid. Dichter bij de waarheid komen, brengt de gehele mensheid vooruit want het zorgt voor meer kennis. Er kan hieruit geconcludeerd worden dat hoe vreemd een opinie ook klinkt; de kans dat er een gedeelte van de waarheid in zit, is al genoeg reden dat de opinie serieus genomen dient te worden. Vrijheid van meningsuiting staat volgens Mill dus op gespannen voet met waarheidsvinding, integendeel, vrijheid van meningsuiting is nodig precies om aan waarheidsvinding te kunnen doen.

5.2. Mills argumenten in het digitale tijdperk: vrijheid van meningsuiting, waarheidsvinding en censuur

Mills visie op vrijheid van meningsuiting is nog steeds relevant; door meer mensen de kans te geven zich te uiten, zijn we als mensheid ook vooruit gegaan op meerdere culturele en technologische aspecten (waaronder ons internet). Het is dan ook niet gek dat men regelmatig nog teruggrijpt naar zijn tekst wanneer er sprake is van censuur.

Het is interessant om te bedenken hoe Mill zelf zijn eigen boek tegen het licht van ons digitale tijdperk zou houden wanneer hij nu, anno 2021, zou leven. Er kunnen wel degelijk

antwoorden worden gevonden in zijn tekst die relevant zijn voor het digitale tijdperk in de 21e eeuw. Zo schrijft hij dat het een plicht is van de overheid maar ook van individuen om zorgvuldig de eerlijkste vorm van opinie te formuleren voordat het opgelegd mag worden aan anderen in een debat (Mill, 2009: 33). Denk hierbij aan criteria zoals hoor- en wederhoor en het 'factchecken' van informatie die als belangrijk worden geacht in de hedendaagse journalistiek. Daarnaast mag men volgens Mill nooit gedwongen of bedrogen worden bij het overtuigen van anderen onder het mom van 'vrijheid zich te verenigen' (Mill, 2009: 23). Ten slotte stelt Mill dat zelfs wanneer iemands uitspraken niet direct iemand raken, dan nog moet hij of zij gedwongen worden zich te beheersen ter wille van hen die door het zien van de informatie misleid worden (Mill, 2009: 136).

Er zijn twee dingen die we kunnen vaststellen aan de hand van Mills argumenten over *information disorders* op sociale media platformen. Ten eerste, wanneer we Mills lijn van denken volgen dan is het wel degelijk gerechtvaardigd om *information disorders* te censureren wanneer ze schade toebrengen. Als een digitaal platform zoals Facebook dreigt schade toe te brengen of het platform faciliteert dat X schade aan Y toebrengt, dan heb je een basis voor dwingende overheidsingrijpen, zoals censuur. Wanneer *information disorders* geen schade toebrengen, dan wordt het wat complexer. Feitelijk onjuiste informatie kan ons simpelweg niet vooruithelpen richting de waarheid en dat is kwalijk, maar vrijheid van meningsuiting helpt ons bij een groter proces van waarheidsvinding. Dit grotere proces houdt in dat *information disorders* die geen schade toebrengen, uiteindelijk zelfstandig uit het publieke debat worden gefilterd (vanuit Habermas en Mills perspectief). Echter kunnen we ons steeds meer afvragen of dit proces nog wel plaats kan vinden via sociale media platformen.

Ten tweede, zijn sociale media platformen niet (meer) op de juiste manier ingericht om Mills gewenste 'marktplaats vol ideeën' te representeren. Sociale media platformen zijn eerder getransformeerd tot een 'Poolse landdag': een vergadering die rommelig verloopt met veel mensen die alleen maar door elkaar praten. De publieke sfeer op sociale media is namelijk rommelig en extremen eisen veelal de aandacht op. Mill (2009) zag altijd hoop in het idee dat mensen naar verschillende opinies luisteren (87), maar door de vormgeving en algoritmes die worden ingezet door digitale media is er steeds minder sprake van deze hoop. Met andere woorden, vrijheid van meningsuiting is belangrijk in het kader van een liberale samenleving waar vrije uitwisseling van informatie meerwaarde heeft voor de wetenschap en onze democratie. Die meerwaarde wordt echter op dit moment niet volledig gerealiseerd door sociale media platformen. Hoe kunnen we de meerwaarde van vrije uitwisseling dan wél realiseren?

En, hoe bestrijden we het beste de *information disorders* die zich verspreiden op deze platformen?

5.3. Hoe nu verder?

In de vorige hoofdstukken heb ik beargumenteerd waarom sociale media platformen een goede broeiplaats zijn voor *information disorders* en waarom deze platformen omtrent het principe 'waarheidsvinding' een morele verantwoordelijkheid hebben. Daarnaast heb ik aan de hand van Mill beargumenteerd dat het principe vrijheid van meningsuiting niet op gespannen voet staat met het principe waarheidsvinding, maar dat het juist een middel tot het doel is. We hebben vrijheid van meningsuiting nodig, maar *information disorders* hebben geen recht van bestaan volgens het principe van vrijheid van meningsuiting wanneer ze schade toebrengen aan anderen.

Sociale media platformen zouden bij uitstek een handje kunnen helpen als het gaat om waarheidsvinding, want: het zou dé plek kunnen zijn waar een pluraliteit aan informatie en opinies te vinden is. Dat pakte echter even anders uit. De onlinewereld is eerder een fragmentatie van de publieke sfeer geworden zoals reeds besproken is in de andere paragrafen. Toch hebben sociale media platformen een belangrijk rol in de publieke sfeer en die trend gaat hoogstwaarschijnlijk niet veranderen. Des te belangrijker is het om na te blijven denken over hoe we die publieke sfeer wél zo goed mogelijk vorm kunnen geven online met als uiteindelijke doel waarheidsvinding. Er is een mismatch tussen wat wij als liberale samenleving willen (namelijk waarheidsvinding) en wat er op dit moment op sociale media platformen plaatsvindt. Het is dus belangrijk dat deze platformen waarheidsvinding in ieder geval niet in de weg zitten.

Er zijn twee oplossingen waardoor sociale media platformen – de voor een goed functionerende democratie belangrijke - waarheidsvinding niet meer in de weg zitten: (1) ze kunnen uit de nieuwssector stappen en enkel entertainment blijven aanbieden en (2) als ze wel in de nieuwssector willen blijven, moeten ze zich aan bepaalde design principes houden die meer in het teken staan van waarheidsvinding dan van hun private winstmodel. Bij de tweede optie zou het al een startpunt zijn wanneer sociale media platformen hun gebruikers niet meer zien als slechts consumenten maar ook als democratische burgers. Dit geeft namelijk ruimte om beter na te kunnen denken over een vormgeving die in het teken staat van waarheidsvinding. Van een traditioneel nieuwsmedium verwachten we immers ook dat ze zich aan principes houden die waarheidsvinding stimuleren met het oog op een goedwerkende democratie.

Daarnaast kan het idee van een deliberatieve democratie - om waarheidsvinding te faciliteren - goed werken wanneer sociale media platformen hun platformen daar actiever op zouden inrichten.

Denk hierbij aan het aanpassen van de algoritmes (zoals besproken in 3.3) of aan het creëren van tools waarbij mensen makkelijker op een respectvolle manier in gesprek kunnen gaan met mensen die een andere opinie hebben[5]. Daarnaast zouden de digitale platforms kunnen overwegen om juist actief gebruik te maken van het menselijk redeneren en de 'wijsheid van de massa'. Zo heeft recent onderzoek aangetoond dat wanneer je mensen online informatie kort laat evalueren, de meerderheid *information disorders* er goed weet uit te pikken. Op deze manier kunnen platformen crowdsourcing inzetten om *information disorders* op te sporen (Pennycook & Rand, 2021: 397). En nog veel belangrijker, wanneer mensen even kort na moeten denken over de accuraatheid van informatie, helpt het de gebruikers van de platformen om te stoppen met het verspreiden van *information disorders* zoals we hebben besproken in 3.2 bij het Harvard onderzoek (Fazio, 2020). Met andere woorden, omdat het probleem van de verspreiding van *information disorders* vaak te maken heeft met onoplettendheid (automatische processen van systeem 1) en het inspelen op allerlei psychologische biases, is het een succesvolle aanpak om het design van de platformen zo in te richten dat het mensen juist ertoe zet om gebruik te maken van hun redeneer skills. Op die manier 'nudge' je mensen in het nadenken over de validiteit en accuraatheid van de informatie die ze tot consumeren.

Zoals besproken in de introductie van deze scriptie, zien veel sociale media platformen een oplossing in het censureren van *information disorders*. Dit is echter dweilen met de kraan open: het alleen verwijderen en labelen van *information disorders* is niet genoeg wanneer hun eigen business model en design medeverantwoordelijk is voor het verspreiden van deze informatie. Wanneer we het probleem echt willen aanpakken, dan zullen we moeten werken aan de bron van het probleem (en dus het design van het platform) en niet het design behouden om vervolgens de (voorspelbare) problemen te moeten blijven opruimen die voortvloeien uit dit design.

Als vervolgens dan blijkt dat sociale media platformen deze verantwoordelijkheid niet nemen, dan is regulatie vanuit de overheid essentieel. Wanneer bedrijven veel impact op de samenleving hebben, dan vinden we het over het algemeen noodzakelijk dat er gereguleerd

[5] Er zijn al een aantal (nog relatief kleine) platformen gestart die als doel hebben om mensen met elkaar te laten debatteren over diverse onderwerpen. Voorbeelden hiervan zijn: Create Debate, Kialo, Opposing Views en Debate Map.

wordt zodat de impact niet te gevaarlijk wordt. Het is daarom nodig dat sociale media platformen in zekere mate beter gereguleerd worden door de overheid om te zorgen dat het maatschappelijk belang van goed geïnformeerde mensen niet lijdt onder financieel gewin. Daarnaast kunnen we er vanuit gaan dat het aanpassen van het design wel veel problemen wegneemt, maar dat er nog steeds problemen kunnen blijven ontstaan. Het is dan belangrijk dat er een wettelijk kader komt vanuit de overheid. Zo zijn bijvoorbeeld het oproepen tot haat en het aanzetten tot geweld onderwerpen waarbij de verantwoordelijkheid niet alleen bij de bedrijven gelegd kan worden, maar ook bij de overheid. Zoals we hebben gezien bij Mill is vrijheid van meningsuiting niet absoluut en zijn bovenstaande onderwerpen een legitieme basis voor een overheid om berichten te censureren en er zelfs straffen op te leggen[6].

Er kunnen natuurlijk ook vragen gesteld worden bij de (morele) verantwoordelijkheden van de gebruikers van sociale media platformen: in hoeverre zijn zij niet ook zelf verantwoordelijk voor de verspreiding en het in stand houden van *information disorders*? Dit is een vraagstuk dat ik niet behandel in deze scriptie maar het is wel een interessant onderwerp voor verder onderzoek. Echter blijf ik er bij dat we eerst oplossingen moeten zoeken bij de bron (de platformen zelf) van het probleem, voordat we ons kunnen storten op de bovenstaande vraag. Wanneer je een snelweg maakt vol met gevaarlijke bochten, dan moet je het namelijk niet gek vinden dat de gebruikers van de weg er samen een puinhoop van maken. Dus zelfs als de gebruikers een deel van de verantwoordelijkheid dragen voor de (dis)informatie die ze delen, dan nog dragen de platformen de verantwoordelijkheid in de mate dat ze dit 'triggeren' en faciliteren door hun doelbewust design.

Kortom, de oplossing zit helemaal niet in het blijven opzoeken van het spanningsveld tussen censuur en vrijheid van meningsuiting. De oplossing zit in het aanpakken van de bron: het design van sociale media platformen. Op deze manier wordt censuur namelijk vanzelf minder relevant omdat er uiteindelijk minder makkelijk *information disorders* verspreid zullen worden.

[6] Zodoende bestaat er sinds 2016 'The EU Code of Conduct on countering illegal hate speech online' waar alle bekende platformen, zoals Instagram en Facebook, bij aangesloten zijn. Hierin staan afspraken over hoe er om gegaan moet worden met situaties waarin er sprake is van haatdragende content (European commission, 2021).

6. Conclusie

In deze scriptie onderzocht ik de vraag: ‘wat is de morele verantwoordelijkheid van sociale media platformen betreffende het verspreiden van *information disorders*, in het licht van vrijheid van meningsuiting?’.

Ik beargumenteer dat de combinatie van het business model en het design van de bedrijven achter de sociale media platformen met onze psychologische neigingen, waar deze bedrijven op inspelen, een broeiplaats vormen voor *information disorders*. De digitale platformen zijn daarom medeverantwoordelijk voor de verspreiding van deze informatie. Sociale media platformen zoals Facebook, Twitter en YouTube hebben bovendien een *morele* verantwoordelijkheid om actief *information disorders* te bestrijden. Aan de hand van vijf voorwaarden (capaciteit, causaliteit, kennis, vrijheid en kwaad-doen) heb ik beargumenteerd dat sociale media platformen morele actoren zijn die, door hun grote impact op de samenleving, moreel verantwoordelijk gehouden kunnen worden. Vervolgens heb ik beargumenteerd dat sociale media platformen in hun huidige vorm het principe ‘waarheidsvinding’ ondermijnen. Dit terwijl ze juist goed zouden kunnen bijdragen aan het vergroten van waarheidsvinding: een deliberatieve democratie (discussies tussen geïnformeerde individuen leiden uiteindelijk tot een zekere vorm van consensus) zou mogelijk kunnen zijn via sociale media platformen. Ten slotte heb ik aan de hand van Mill beargumenteerd wat de rol van het principe ‘vrijheid van meningsuiting’ is binnen het bestrijden van *information disorders*.

Vrijheid van meningsuiting is een middel dat we gebruiken bij ons streven naar waarheidsvinding. *Information disorders* vallen niet geheel onder vrijheid van meningsuiting omdat ze (1) ons helemaal niet vooruit helpen richting de waarheid (het is namelijk feitelijk onjuist) en (2) het mensen schaadt (voorbeelden zijn besproken in paragraaf 2.3). Mill en ook Habermas geloofden dat *information disorders* niet de boventoon zouden voeren wanneer we als goed geïnformeerde mensen een fatsoenlijk en redelijk debat konden voeren. Sociale media platformen falen op dit moment echter in het faciliteren van deze digitale democratische plek. De pogingen van sociale media platformen om *information disorders* te bestrijden door ze te censureren, is als dweilen met de kraan open: het alleen verwijderen van *information disorders* is niet genoeg wanneer hun eigen business model en design medeverantwoordelijk zijn voor het verspreiden van de informatie. Er zijn dus hervormingen nodig in het business model en het design van sociale media platformen willen ze echt actief *information disorders* bestrijden. Een andere oplossing is dat deze online platformen stoppen met zich te profileren als nieuwsmedium

en zich (weer) enkel richten op entertainment. Wanneer sociale media platformen deze morele verantwoordelijkheid niet op zich nemen, is het noodzakelijk dat een overheidsinstantie ingrijpt. Het maatschappelijk belang van een open en redelijk publiek debat onder goed geïnformeerde mensen in onze liberale, democratische samenleving mag nooit lijden onder het financieel gewin van de grote digitale platformen.

7. Literatuur

Anderson, I. A., & Wood, W. (2021). Habits and the electronic herd: The psychology behind social media's successes and failures. Consumer Psychology Review, 4(1), 83-99.

Bodner, G.E., Masson, M.E.J. & Richard (2006). Repetition proportion biases masked priming of lexical decisions. Memory & Cognition 34, 1298–1311.

Brady, W. J., Wills, J. A., Burkart, D., Jost, J. T., & Van Bavel, J. J. (2019). An ideological asymmetry in the diffusion of moralized content on social media among political leaders. Journal of Experimental Psychology: General, 148(10), 1802.

Bryant, L.V. (2020). The YouTube Algorithm and the Alt-Right Filter Bubble. Berlin: De Gruyter Open.

Buekens, F. (2018). Woord & Wereld: een inleiding tot de taalfilosofie. Leuven: Acco.

Chitra, U., & Musco, C. (2020). Analyzing the impact of filter bubbles on social network polarization. In Proceedings of the 13th International Conference on Web Search and Data Mining (pp. 115-123).

Darnton, R. (2017). The True History of Fake News. Published in The New York Review.

Davis, K. (1960). Can business afford to ignore social responsibilities?. California management review, 2(3), 70-76.

Dretske, F. I. (1983). Précis of Knowledge and the Flow of Information. Behavioral and Brain Sciences, 6(1), 55-90.

EenVandaag. (2015, 7 oktober). Protest tegen komst AZC in Purmerend. Geraadpleegd via:http://binnenland.eenvandaag.nl/tvitems/vluchtelingen/62393/protest_tegen_komst_azc_in_purmerend

European Commision (2021, 9 juni). The EU Code of conduct on countering illegal hate speech online. Geraadpleegd via: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

Fazio, L. K. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. Harvard Kennedy School (HKS) Misinformation Review.

Fetzer, J. H. (2004). Information: Does it have to be true?. Minds and Machines, 14(2), 223-229.

Floridi, L. (2011). The philosophy of information. Oxford: Oxford University Press.

Garriga, E., & Melé, D. (2004). Corporate social responsibility theories: Mapping the territory. Journal of business ethics, 53(1), 51-71.

Ghebreyesus, T.A. (2020, 15 February). Munich Security Conference. World Health Organization. Geraadpleegd via: https://www.who.int/director-general/speeches/detail/munich-security-conference

González, E. (2002). Defining a post-conventional corporate moral responsibility. Journal of business ethics, 39(1), 101-108.

Grömping, M. (2014). 'Echo Chambers' Partisan Facebook Groups during the 2014 Thai Election. Asia Pacific Media Educator, 24(1), 39-59.

Habermas, J. (2006). Political communication in media society: Does democracy still enjoy an epistemic dimension? The impact of normative theory on empirical research. Communication theory, 16(4), 411-426.

Habermas, J., Lennox, S., & Lennox, F. (1974). The public sphere: An encyclopedia article (1964). New German Critique, (3), 49-55.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological review, 108(4), 814.

O'Hara, K., & Stevens, D. (2015). Echo chambers and online radicalism: Assessing the Internet's complicity in violent extremism. Policy & Internet, 7(4), 401-422.

Informatie. (z.d.). In Van Dale. Van Dale uitgevers. Geraadpleegd op 21 mei 2021, van https://www.vandale.nl/gratis-woordenboek/nederlands/betekenis/informatie#.YLs4zKgzZPY

Kahneman, D. (2011). Thinking, fast and slow. London: Penguin Psychology

Lazer, D., et al. (2018). The Science of Fake News. Science: 359(6380):1094-1096.

Lewis, R. (2018). Alternative influence: Broadcasting the Reactionary Right on YouTube. Data & Society Research Institute.

Mayraz, G. (2011). Wishful thinking. SSRN Electronic Journal.

Meeker, M. (2019). The Internet Trend Report. Geraadpleegd via: https://www.bondcap.com/report/itr19/#view/12

Melé, D. (2008). Corporate social responsibility theories. The Oxford handbook of corporate social responsibility, 47-82.

Mill, J.S. (1859 [2009]). On Liberty. Auckland: The Floating Press.

Moravec, P., Minas, R., & Dennis, A. R. (2018). Fake news on social media: People believe what they want to believe when it makes no sense at all. Kelley School of Business Research Paper.

Musch, S. (2019, 15 juni). EU: individuele Russische trollen probeerden verkiezingen te manipuleren. NRC. Geraadpleegd via: https://www.nrc.nl/nieuws/2019/06/15/russische-trollen-probeerden-europese-verkiezingen-te-beinvloeden-a3963844

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of general psychology, 2(2), 175-220.

NOS. (2020, 24 december). LinkedIn verwijdert account Van Haga vanwege 'verspreiding valse informatie'. Geraadpleegd via https://nos.nl/artikel/2361923-linkedin-verwijdert-account-van-haga-vanwege-verspreiding-valse-informatie.html

Obear, J. (2018). Move Last and Take Things: Facebook and Predatory Copying. Columbia Business Law Review, 994.

Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. London: Penguin UK.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. Trends in cognitive sciences.

Quattrociocchi, W., Scala, A., & Sunstein, C. R. (2016). Echo chambers on Facebook.

Steup, M. & Ram N., "Epistemology", The Stanford Encyclopedia of Philosophy (Fall 2020 Edition). geraadpleegd via: https://plato.stanford.edu/archives/fall2020/entries/epistemology/.

Stevenson, A. (2018). Facebook Admits It Was Used to Incite Violence in Myanmar. The New York Times. Geraadpleegd via: https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html

Terranova, T. (2012). Attention, economy and the brain. Culture Machine, 13.

Tokmetzis, D., Bahara, H., Kranenberg, A. (2019, 8 februari). Aanbevolen voor jou op YouTube: racisme, vrouwenhaat en antisemitisme. De Correspondent. Geraadpleegd via: https://decorrespondent.nl/9149/aanbevolen-voor-jou-op-youtube-racisme-vrouwenhaat-en-antisemitisme/1201966498733-5cf61b02

United Nations. (z.d.). Universal Declaration of Human Rights. Geraadpleegd via: https://www.un.org/en/about-us/universal-declaration-of-human-rights

Van de Poel, I., Royakkers, L., & Zwart, S. D. (2015). Moral responsibility and the problem of many hands. London: Routledge.

Visser, M. (2020, 21 augustus). Vijftig twittertrollen verspreidden bewust nepnieuws over corona. Trouw. Geraadpleegd via: https://www.trouw.nl/binnenland/vijftig-twittertrollen-verspreidden-bewust-nepnieuws-over-corona~b2ed89b4/

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.

Wardle, C. (2017). INFORMATION DISORDER: Toward an interdisciplinary framework for research and policy making. Strasbourg: Council of Europe.

Weisberg, J. (2011). Bubble trouble: Is web personalization turning us into solipsistic twits. Slate. com, 10-06. Geraadpleegd via: https://slate.com/news-and-politics/2011/06/eli-pariser-s-the-filter-bubble-is-web-personalization-turning-us-into-solipsistic-twits.html

Wiener, N. (1954). The human use of human beings. Cybernetics and society. New York: Doubleday Anchor Books.

Wijkers, M. (2021, 9 februari). De belangrijkste cijfers van ons social media gebruik in 2020. Afix. Geraadpleegd via: https://www.afix.nl/belangrijkste-cijfers-social-media-gebruik-2020/#:~:text=96%25%20van%20de%20Nederlanders%20is,minste%20tijd%20op%20social%20media

Williams, J. (2018). Stand out of our light: freedom and resistance in the attention economy. Cambridge: Cambridge University Press.

Wong, C.K. (2020). Facebook restricts more than 10,000 QAnon and US militia groups. The Guardian. Geraadpleegd via https://www.theguardian.com/us-news/2020/aug/19/facebook-qanon-us-militia-groups-restrictions

Wright, S., & Street, J. (2007). Democracy, deliberation and design: the case of online discussion forums. New media & society, 9(5), 849-869.

YouTube Team (2019). Continuing our work to improve recommendations on YouTube. Geplaatst op 25 januari 2019. Geraadpleegd via: https://blog.youtube/news-and-events/continuing-our-work-to-improve

Zuckerberg, M. (2021, 7 januari). Facebook Post. Geraadpleegd via https://www.facebook.com/zuck/posts/10112681480907401

Zuiderveen Borgesius, F. J. & Trilling, D. & Möller, J. & Bodó, B. & de Vreese, C. H. & Helberger, N. (2016). Should we worry about filter bubbles? Internet Policy Review, 5(1).

Zulli. D. (2018) Capitalizing on the look: insights into the glance, attention economy, and Instagram. Critical Studies in Media Communication, 35:2, 137-150.