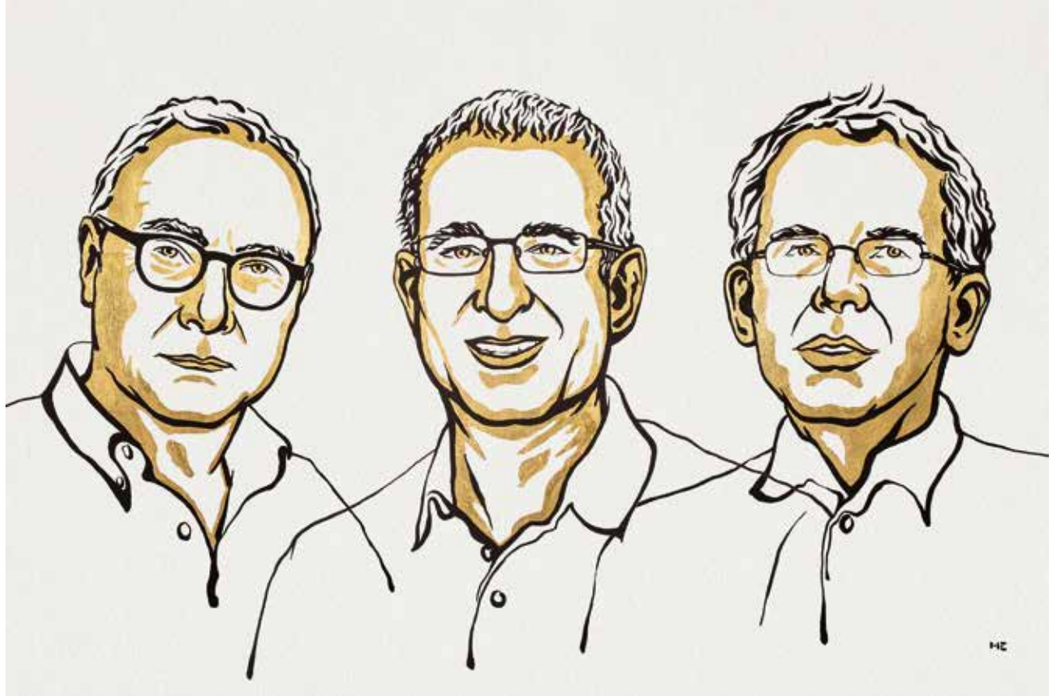


David Card, Joshua Angrist en Guido Imbens, winnaars van de Nobel Economics Prize 2021. Illustratie: Niklas Elmehed | Nobel Prize Outreach

DE REIKWIJDTE VAN DE COUNTERFACTUAL

over causaliteit, potential outcomes en grafische modellen

Richard Starmans

In 2011 ontving de Amerikaanse informaticus Judea Pearl (1936) de door de Association for Computing Machinery (ACM) ingestelde A.M. Turing Award voor zijn fundamentele bijdragen aan de AI *'through the development of a calculus for probabilistic and causal reasoning'*. In 2021 werd aan de van oorsprong Nederlandse econometrist Guido Imbens (1963) en de Amerikaans-Israëlische econoom Joshua Angrist (1960) de Nobel Memorial Price in Economic Sciences toegekend *'for their methodological contributions to the analysis of causal relationships'*. Overigens sleepten Imbens en Angrist samen 'slechts' de helft van de prijs in de wacht, de andere helft viel ten deel aan de Canadese econoom David Card (1956) voor zijn empirische bijdragen aan de arbeidseconomie en de ontwikkeling en toepassing van *natural experiments*. Dat laatste begrip vormt een belangrijke verbindende schakel tussen de drie winnaars. Volgens de jury leverden de laureaten te onderscheiden, maar complementaire bijdragen, waarin wordt gedemonstreerd hoe oorzaak-gevolg relaties kunnen worden vastgesteld en geanalyseerd in natuurlijke experimenten, die betrekking hebben op complexe *real-life problems* en daarmee verbonden beleidsvraagstukken met een grote maatschappelijke impact.

Causaliteit

Het gegeven dat prestigieuze wetenschappelijke prijzen zoals de Nobelprijs voor Economie en de Turing Award worden toegekend aan onderzoek naar *causal inference* of *causal reasoning* is saillant, aangezien causaliteit van oudsher geldt als een obscuur en weerbarstig begrip. Het denken over oorzaak-gevolg relaties kent een lange en moeizame genealogie, die teruggaat tot Aristoteles en de Stoa en door sommigen – terecht of onterecht – louter met metafysica werd geassocieerd en niet met een wetenschappelijk wereldbeeld. Het gaf volop aanleiding tot controversen, kende geen bruikbare formaliseringen en leidde nauwelijks tot vooruitgang. De ideeëngeschiedenis telt dan ook vele prominente denkers die ex cathedra verordonneerden dat in het wetenschappelijk discours geen plaats behoort te zijn voor zulk een archaïsche notie. Men denke aan Ernst Mach, Bertrand Russell, Karl Pearson, Ludwig Wittgenstein of Paul en Patricia Churchland. Ook in tijden van big data, data science en vooral *deep learning* weerklinkt met name in de populaire literatuur de opvatting dat het begrip causaliteit toch in het gunstigste geval als obsoleet dient te worden beschouwd. Onder meer voormalig Google-onderzoeksdirecteur Peter Norvig, econoom Victor Mayer-Schönberger, wetenschapsjournalist Chris Anderson en 'The Master Algorithm'-auteur Pedro Domingos droegen bij aan een rijk geschakeerd palet aan stellingnames, dikwijls gelardeerd met anti-causalistische allusies, waarvan sommige uiteraard genuanceerder en gematigder zijn dan andere.

De toekenning van voornoemde prijzen mag dan aantonen dat vooruitgang op het gebied van causaliteit wel degelijk mogelijk is, de vernieuwde of voortgezette belangstelling ervoor leidt allerminst tot een verenigd en eensgezind veld. Veeleer is er sprake van een veelstromenland, dat een verzuilde aanblik biedt met tal van tegenstellingen en naast elkaar bestaande benaderingen, die (nog steeds) weinig interactie vertonen, zeker in de overvloedige filosofische literatuur (Illari, 2014) (Starmans, 2018; 2020). Dat laatste is wellicht niet verrassend voor diegenen die menen dat in de filosofie nu eenmaal alleen 'vooruitgang' mogelijk is door het plegen van een intellectuele vadermoord, waarbij radicaal wordt gebroken met de traditie waaruit men voorkomt. De tegenstellingen blijken evenwel ook manifest als we ons beperken tot moderne probabilistische benaderingen, waarin de geschetste vooruitgang nu juist werd geboekt en die vooral in de statistiek, de AI, econometrie en deels in de sociale wetenschappen opgang hebben gemaakt. Opmerkelijk genoeg staan ook de voornoemde laureaten Imbens en Pearl in een aantal opzichten tegenover elkaar en bekennen zich tot twee causale tradities, waarvan de verschillen nogal eens worden uitvergroot (Pearl, 2018; 2021; Imbens, 2020). Waar Pearl vooral binnen de AI furore maakte met zijn grafische modellen of *directed acyclic graphs* (DAG), zijn *do-calculus* en zijn *backdoor-* en *frontdoor* criteria voor identificatie van causale effecten, werkt Imbens vooral binnen de economie aan de integratie van de methode der instrumentele variabelen met de zogenaamde Potential Outcome benadering (PO). Enkele aspecten van de problematiek brengen we hier kort voor het voetlicht.

Counterfactuals

De tegenstelling tussen de PO- en DAG-benadering is opmerkelijk, omdat de verbondenheid verder gaat dan louter de constatering dat beide benaderingen probabilistisch zijn en claimen dat causale verbanden wel degelijk in observationele data kunnen worden vastgesteld. Allereerst zijn de methoden wiskundig grotendeels equivalent; een stelling binnen het DAG-raamwerk is dan een stelling binnen de PO-benadering en vice versa. Daarnaast komen beide voort uit twee statistische en sterk causaal-georiënteerde tradities, die teruggaan tot de vroege jaren twintig van de vorige eeuw. De DAG's vinden hun wortels in de padmodellen van Sewall Wright, een causalist van het eerste uur aan wie Pearl zich in (Pearl, 2018) schatplichtig toont en die hij roemt als herauut en pionier van de door hem geproclameerde causale revolutie (Wright, 1921). De PO-benadering is terug te voeren tot het vroege werk van Jerzy Neyman, die de methode introduceerde in het kader van inzichten van Ronald Fisher op het terrein van *experimental design* en inferentiële statistiek (Neyman, 1923). Later zou Donald Rubin de methode verder ontwikkelen in een context van observationele data. De PO-aanpak wordt dan ook dikwijls getypeerd als Neyman-Rubin modellen. Daarbij mag ook de econometrische invalshoek niet worden vergeten, met name de simultaneous equations methode van Jan Tinbergen, Philip Wright en later Trygve Haavelmo. Philip Wright, de vader van Sewall Wright, wees als een van de eersten op het concept van instrumentele variabelen bij het oplossen van het identificatieprobleem bij vraag-aanbod modellen.

De voor dit korte essay meest relevante overeenkomst tussen DAG's en PO is echter filosofisch en betreft het gebruik van *counterfactuals* als sleutel tot causaliteit. Zo beschouwd impliceert causaal redeneren nadenken over, verwijzen naar, postuleren van en uitspraken doen over tegenfeitelijke werelden, parallel aan de feitelijke of actuele wereld; uitspraken waaraan wel een waarheidswaarde in die tegenfeitelijke wereld lijkt te kunnen of zelfs moet worden toegekend vanuit de actuele wereld. Counterfactuals verwijzen naar verschillende klassen van uitspra-

ken, die we hier gemakshalve typeren als uitspraken van de vorm: als P het geval zou zijn / het geval was geweest / had plaatsgevonden, dan zou Q het geval zijn / het geval zijn geweest / hebben plaatsgevonden. Uiteraard kunnen antecedens, consequens of beide de negaties van P en Q bevatten. Essentieel is hierbij dat P kan verwijzen naar een observatie, feit, toestand, handeling of gebeurtenis en dat P in de actuele wereld verondersteld wordt niet waar te zijn, niet te hebben plaatsgevonden, maar in de tegenfeitelijke wereld wel en vice versa. Ter illustratie wordt vaak het onderscheid gemaakt tussen indicatieve conditionals (als P het geval is, dan is ook Q het geval) en subjunctive conditionals (Als P het geval zou zijn, dan zou ook Q het geval zijn), dat iets van de problematiek verheldert met de toevoeging dat elke counterfactual een subjunctieve conditional is, maar niet omgekeerd. In de algemene vorm van de subjunctieve conditional kan de spreker agnostisch zijn over P, zonder een sterke claim over een tegenfeitelijke wereld, terwijl dit bij een counterfactual wel het geval is (Starmans, 2021a).

Hoe dan ook, counterfactuals werden al vroeg verbonden met de studie van causaliteit en het was de sceptische filosoof David Hume (1711–1776) die daarbij het voortouw nam. Zowel zijn vroege en lijvige *A Treatise of Human Nature* uit 1739 als zijn latere, meer toegankelijke en populaire *An Enquiry concerning Human Understanding* uit 1748 bevat in dit opzicht cruciale passages. In Sectie VII van de *Enquiry* stelt de auteur onder meer: '*We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or, in other words, where, if the first object had not been, the second never had existed*.' Met name de frase 'in other words' is frappant en verbindt het eerste gedeelte (fragment A), dat een notie van causaliteit als regulariteit of constante conjunctie c.q. opeenvolging in de tijd suggereert met het tweede gedeelte (fragment B), dat feitelijk een counterfactual behelst. Duidelijk is dat de auteur hiermee een sterke verbondenheid van beide suggereert; twee typeringen die equivalent zijn, of – iets zwakker wellicht – middels een relatie van logisch gevolg op elkaar betrokken zijn, of die toch zeker een sterke semantische verwantschap vertonen, waarbij fragment B dient ter verduidelijking of verdere afbakening van fragment A. Waarschijnlijk beschouwde Hume zelf de tegenfeitelijke beschrijving als een voor de lezer meer vertrouwde en intuïtieve notie, een commonsense uitleg, die kon helpen de eerste beschrijving te duiden.

Causaliteit en counterfactuals

In ten minste drie opzichten is dit alles opmerkelijk. Allereerst moet worden opgemerkt dat in de daaropvolgende eeuwen de fragmenten A en B zich langs gescheiden paden zouden ontwikkelen en uitgroeien tot zelfstandige benaderingen van causaliteit. Vandaag de dag gelden de regulariteitsbenadering en de counterfactual-benadering als twee afzonderlijke invalshoeken, naast vele andere. Beide zijn noch tot elkaar te herleiden, noch fungeert de een ipso facto als verduidelijking van de ander. Daarbij komt dat vanuit Hume's strikte concept-empirisme, waarin begrippen rechtstreeks tot impressies te herleiden moeten zijn, het speculeren over niet-geobserveerde c.q. niet-observeerbare toestanden of mogelijke werelden twijfelachtig is. Ze mogen dan als vanzelfsprekend opduiken in gedachte-experimenten, die de filosofie van oudsher heeft gekend, maar een strenge empiristische epistemologie doet doorgaans geen beroep op noties die niet-waarneembaar, speculatief, metafysisch of wellicht zelfs incoherent zijn en niet tot sense data of impressies zijn te herleiden. Zij kunnen hoogstens een illustratieve of heuristische functie vervullen en dienen met de nodige scepsis te worden benaderd. In de derde plaats is het opmerkelijk dat Hume toen reeds de counterfactual te hulp riep om de causale relatie te duiden, omdat de counterfactual een notoir lastig begrip was en in zekere zin nog steeds is. Lange tijd golden counterfactuals als obscuur, een precieze interpretatie bleek problematisch en pogingen tot een waarheidsfunctionele semantiek te komen waren voor velen niet overtuigend. Het zou duren tot ver in de twintigste eeuw voordat men enige greep kreeg op counterfactuals, onder meer door de logisch-linguïstische benadering van Saul Kripke (1940) en zijn mogelijke werelden semantiek, maar vooral met het werk van David Lewis (1941–2001), wiens Counterfactuals uit 1973 inmiddels de status van een moderne klassieker heeft verworven. Maar nog steeds bestaat over de reikwijdte ervan geen volledige consensus. Een traditioneel bezwaar betreft het feit dat counteractuals niet waarheidsfunctioneel zijn: de waarheidswaarde van het geheel is geen functie van de waarheidswaarde van de delen, de deelpremissen. Daar komt bij dat counterfactuals in verschillende talen verschillend worden gerepresenteerd, vooral met betrekking tot tense, modaliteit en aspect. Dat geldt zeker voor het gebruik van de zogenaamde *fake-tense* om de tegenfeitelijke wereld te evoceren. De notie roept dan ook tal van filosofische vragen op: *ontologisch* (kan zo'n wereld bestaan en hoe dan?), *epistemologisch* (hoe kunnen we kennis verwerven van deze wereld, die we niet waarnemen en waarin we niet kunnen tellen, meten en wegen en evenmin interveniëren?), *semantisch* (hoe kunnen we betekenis toekennen aan uitspraken over die wereld vanuit de actuele wereld?) en *pragmatisch* (Is het gebruik ervan eigenlijk wel nodig of wenselijk, zeker in een wetenschappelijke context?). Vele aspecten blijven hier buiten beschouwing, maar de vraag is uiteraard in hoeverre het mogelijk is in een gedachte-experiment wijzigingen in de actuele wereld te bedenken, deze vervolgens te projecteren op een tegenfeitelijke wereld, de (logische en fysische) consequenties hiervan te kennen, vervolgens de coherentie of bestaanbaarheid ervan te postuleren om zo tot een zinvolle vergelijking tussen beide werelden te komen. Men neme het volgende voorbeeld:

> Als Julius Caesar een kat of een priemgetal was geweest, dan zou de wereldgeschiedenis nu een keizer/veldheer hebben gehad die goed kon klimmen of een persoon die bij de rivier de Rubicon de woorden veni, vidi, vici uitsprak en die enkel deelbaar is door 1 of door zichzelf.

Dit voorbeeld mag artificieel zijn en lijkt tot absurditeiten te voeren, waarbij alle klassieke taalfilosofische problemen betreffende *rigid designators*, *crossworld-identity* en consistentie, overerving van eigenschappen, presupposities, ceteris paribus clausules, et cetera opduiken. De kernvraag is uiteraard welke beperkingen dan blijkbaar moeten worden opgelegd aan de menselijke fantasie en zijn ongebreideld vermogen tot imaginatie en aan zijn taalgebruik om daarmee een vorm van hypothetisch redeneren af te dwingen, waarbij deze tegenfeitelijke wereld wel zinvol gedacht kan worden en er, al dan niet in de vorm van een counterfactual, betekenisvolle uitspraken over kunnen worden gedaan. Welke toestanden of variabelen kunnen in zo'n gedachte-experiment redelijkerwijs gewijzigd worden en welke interacties treden op? Kortom: hoe moet de counterfactual beteugeld worden? (Starmans, 2021b)

PO of DAG?

Al met al kan worden opgemerkt dat Hume in zekere zin een vooruitziende blik had, omdat de counterfactual sedert de late 20e eeuw een ware zegetocht beleeft in causale probabilistische formalismen. Daartoe moesten wel de nodige obstakels worden overwonnen en het succes schuilt deels in de wijze waarop de counterfactual aan banden wordt gelegd. In de PO-aanpak wordt de oorzaak-gevolg relatie geanalyseerd door aan elke onderzoekseenheid een tweetal *potential outcomes* Y(1) en Y(0) toe te kennen, die op een subtiele wijze worden geassocieerd met een gemeten afhankelijke variabele Y. Een causaal effect wordt geschat door de wereld waarin een proefpersoon interventie A ondergaat, $Y_i(1)$, te vergelijken met een wereld waarin diezelfde proefpersoon interventie A niet ondergaat, $Y_i(0)$. Het is uiteraard een fysische onmogelijkheid om in beide werelden tegelijkertijd te vertoeven en metingen te verrichten, hetgeen door sommigen zelfs wordt beschouwd als *The Fundamental Problem of Causal Inference*.

De PO-benadering is derhalve een uitgewerkt gedachte-experiment, waarin dit alles wel mogelijk is en men waarden gaat toekennen aan variabelen c.q. grootheden in zowel de feitelijke of 'actuele' wereld als in een tegenfeitelijke wereld. Men kan dit alles ook opvatten als een missing-data problem, waarbij elke individu twee extra kolommen in de dataset ontvangt, waarvan precies één kolom een missing value heeft, afhankelijk van interventie A. Men kan dan met imputatiemethoden alsnog een ieder twee scores toekennen, een individueel causaal effect $Y_i(1) - Y_i(0)$ berekenen, om daarna bijvoorbeeld voorspellingen voor nieuwe cases te doen, et cetera. Degene voor wie dit een brug te ver is kan ook een causale grootheid definiëren, zoals b.v. een average treatment effect ATE = E [Y(1) – Y(0)], gedefinieerd in termen van potential outcomes, die dan gereduceerd moet worden tot een statistische grootheid, zoals bijvoorbeel *associational difference* E(Y= 1 | A=1) – E(Y=1 |A=0), waarin de potential outcomes zijn geëlimineerd en op basis van geobserveerde data de causale relatie met behulp van conditionele afhankelijkheden wordt geduid. Uiteraard is zulk een reductie niet triviaal. Vanwege lineariteit van de verwachtingsoperator E geldt weliswaar E[Y(1) – Y(0)] = E[(Y1)] – E[(Y0)], maar dit is doorgaans niet gelijk aan E(Y | A=1) – E(Y |A=0). Anders zou het beroemde maxime *causation is not association* onwaar zijn. Zeker als de data niet middels een RCT zijn gegenereerd moet er rekening worden houden met confounding. De associatie tussen confounder Z en interventie A wordt 'gebroken' via gangbare methoden als stratificatie en vooral matching, of

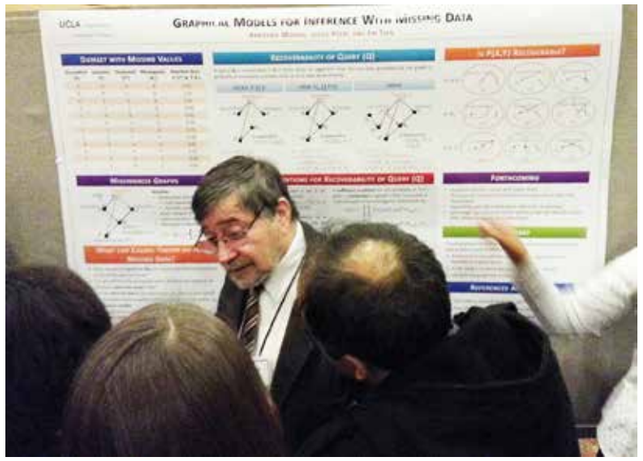
door het opbouwen van een pseudo-populatie middels (*inverse probability*) *weighting*. Een belangrijke aanname om de reductie mogelijk te maken is die van *ignorability* oftewel $(Y(1), Y(0)) \perp\!\!\!\perp A$, waardoor de potential outcomes conditioneel onafhankelijk zijn van A en de *missing values* van de *potential outcomes* genegeerd of *ignored* kunnen worden. Deze aanname wordt ook wel *exchangeability* genoemd, die vrijwel identiek is en inhoudt dat het verwisselen van de *treatment* en *non-treatment* groepen leidt tot dezelfde *potential outcomes*. De groepen zijn dan op de interventie na gelijk en uitwisselbaar. Helaas is deze aanname van *ignorability/exchangeability* in observationele data onrealistisch en men zou daarom kunnen proberen dit in subgroepen van een relevante derde variabele W te bewerkstelligen door te conditioneren. De assumptie van *conditional exchangeability*, oftewel $(Y(1), Y(0)) \perp\!\!\!\perp A \mid W$ leidt dan tot de bescheidener reductie:

$$E[Y(1) - Y(0) \mid W] = E[Y(1) \mid W] - E[Y(0) \mid W] = E(Y \mid A=1, W) - E(Y \mid A=0, W).$$

Het bedoelde marginale effect kan worden bereikt door W 'uit te marginaliseren'.

$$E[Y(1) - Y(0)] = E_W [E[Y(1) - E[Y(0) \mid W]] = E_W [E [Y \mid A=1, W] - E[Y \mid A=0, W]]$$

Deze zeer onvolledige weergave van PO negeert assumpties zoals *positivity* en *unmeasured confounders*, maar illustreert hoe een aloud gedachte-experiment formeel kan worden uitgewerkt en hoe de counterfactual beteugeld kan worden door het streven werelden identiek te houden en alleen te laten verschillen met betrekking tot de interventie. Nog prominenter duikt de counterfactual op in de DAG-benadering van Pearl, die decennialang heeft gewerkt aan een probabilistische benadering die zijns inziens meer recht doet aan de oorzaak-gevolg relatie dan de traditionele statistische benaderingen. Hij spreekt in (Pearl, 2018) zelfs van de 'ladder van causaliteit', waarbij de door hem bepleite verandering in het wetenschappelijke denken middels een drietal stappen of treden gestalte moet krijgen: die van associatie, vervolgens interventie en tot slotte de counterfactual. Hume's regulariteit en de traditionele statistiek blijven volgens de auteur steken op de eerste trede, die van associatie en correlatie, waarbij we alleen kunnen observeren. Bijbehorende vragen als 'Hoe hangen X en Y samen?' en 'Hoe verandert kennis



Judea Pearl tijdens de Conference on Neural Information Processing Systems in 2013

van Y indien X wordt waargenomen?' laten geen diepere, causale conclusies toe. De tweede trede betreft niet alleen waarnemen, maar ook handelen, interveniëren en laat vragen toe als: 'Hoe verandert Y als we actief X veranderen?' De derde trede is volgens Pearl essentieel om echt causaal te kunnen redeneren. Het gaat om *imagining, retrospection* en laat vragen toe als 'Zou Y het geval zijn geweest, als X het geval was geweest?'. Daarmee vormt de counterfactual sterker dan in de PO-aanpak het sluitstuk van de causale redenering en een wezenlijk kenmerk van de menselijke conditie; zonder deze dreigt het gehele project van de sterke AI te mislukken en wordt / blijft de mens overgeleverd aan deep learning. (Pearl, 2018) vormt in feite een lange, soms polemische aanklacht tegen de traditionele statistiek en epistemologie, die volgens hem op de eerste trede van de ladder zijn blijven steken, maar soms ook tegen de methode van Rubin en Imbens, en tegen iedereen die het belang van de grafische representaties niet erkent. Die vormen inderdaad een belangrijk hulpmiddel bij het specificeren van het causale model, identificatie van het effect en het begrijpen van confounding, mediation of effect-modification, maar dit aspect moet hier verder buiten beschouwing blijven.

Imbens lijkt overigens niet geïnteresseerd in de mentalistische claims van Pearl en onderzoekt waarom de DAG's in de economische literatuur nauwelijks voet aan de grond krijgen (Imbens, 2020). Zo is de PO-benadering zijns inziens, ook historisch beschouwd, veel geschikter voor echte, grootschalige (causale) problemen in de economie, terwijl de DAG-benadering veeleer een oplossing lijkt 'op zoek naar' een probleem of vooral tracht *toyproblems* op te lossen. Ook stelt Imbens dat (Pearl, 2018) vooral gaat over identificatie van het causale effect en niet over wat er aan vooraf gaat (model-specificatie, het tekenen van de causale graaf) en wat er na komt (inference!). Bovendien kent het domein van de economie voorwaarden zoals monotoniciteit, die in de DAG-benadering niet goed te formuleren zijn. A fortiori suggereert Imbens dat Pearls ladder met een vierde trede moet worden uitgebreid, die van *reversed causality*, zodat ook van gevolg naar oorzaak kan wordt geredeneerd. Het weerwoord (Pearl, 2021) liegt er evenmin om en het laatste woord over deze kwestie is uiteraard nog niet gezegd. Toch wordt de soep niet altijd zo heet gegeten. Veel moderne literatuur en MOOCS op het gebied van causal inference zijn hybride en eclectisch; concepten en notaties van PO en DAG's worden door elkaar gebruikt en vooral selectief, wanneer het uitkomt en dat geldt met name binnen *machine learning*.

Epiloog

De filosoof Willard V. O. Quine schreef ooit in zijn *Methods of Logic* (1950)*: '... any adequate analysis of the contrafactual conditional* (counterfactual, RS) *must go beyond truth values and consider causal connections, or kindred relationships, between matters spoken of in the antecedent of the conditional and matters spoken of in de consequent*.' Quine, nota bene zelf empirist pur sang, behaviorist en fysicalist stelt dat de counterfactual alleen te begrijpen is vanuit het 'obscure' begrip der causaliteit en niet tot een extensionele logica is te herleiden. Indien causaliteit nodig is om de counterfactual te begrijpen en vice versa dreigt een vicieuze cirkel te ontstaan, al lijkt de recente vooruitgang in het probabilistische causale onderzoek dit te logenstraffen. Toch is het gebruik van causale uitdrukkingen, waarvan ons taalgebruik is doordrenkt, dikwijls kwalitatief, deterministisch en zeker niet probabilistisch. Pogingen het begrip te axiomatiseren indachtig het adagium 'zonder counterfactual geen causaliteit' zijn bovendien problematisch omdat de counterfactual in het dagelijkse taalgebruik moeilijker te beteugelen lijkt. Iemand die in een dialoog/taalspel de uitspraak 'P veroorzaakt Q' doet, daarmee een taalhandeling verricht en bijbehorende commitments aangaat inzake de feitelijke wereld die hij kent, zou zich evenzeer moeten committeren aan uitspraken en een daarmee geassocieerde tegenfeitelijke wereld, die hij niet kent en wellicht niet kenbaar of zelfs incoherent is. Zouden minder weerbarstige 'als–dan'-beweringen niet ten minste overwogen moeten worden? De vraag welke verdedigingsplicht iemand op zich neemt is vooral relevant in tijden van Explainable AI (EAI). (Miller, 2019) beargumenteert dat het zoeken naar (causale) verklaringen zonder een sociale en linguïstische context niet zinvol is en bovendien betreft EAI niet zozeer de context van discovery en de onderliggende causale structuur van de werkelijkheid, maar veeleer de context of justification, waarin oorzaken, redenen en motieven in een gespecificeerd, retorisch gefaseerd taalspel gestalte moeten geven aan de gezochte verklaringen (Starmans, 2020).

Dit alles neemt niet weg dat met de opmars van *causal inference* de counterfactual als aloud gedachte-experiment volop in ere is hersteld.

Literatuur

Illary, P., & Russo, F. (2014). *Causality; philosophical theory meets scientific practice*. Oxford.

Imbens, G. (2020). Potential Outcome and Directed Acyclic Graph approaches to causality; Relevance for empirical practice in economics. *Journal of Economic Literature, 58*(4)

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. *Statistical Science, 5*(4), 1923.

Pearl, J and D. MacKenzie (2018). The Book of Why, the new science of cause and effect, New York.

Pearl, J. (2021); http://causality.cs.ucla.edu/blog/index.php/2020/01/29/on-imbens-comparison-of-two-approaches-to-empirical-economics/

Starmans, R. J. C. M. (2020). Prometheus unbound or Paradise regained: the concept of causality in the contemporary AI-data science debate. *Journal of the French Statistical Society, 161*(1), 4–41.

Starmans, R. J. C. M. (2021a). Over padmodellen, structurele vergelijkingen en de roep om Explainable AI. *STAtOR, 22*(2).

Starmans, R. J. C. M. (2021b). De tegenfeitelijke wereld van de counterfactual; obscuur gedachte-experiment of sleutel tot de causale redenering. *Filosofie Tijdschrift, 31*(5).

Wright, S. (1921). Correlation and Causation. *Journal of Agricultural Research, 20*, 557–585.

Richard Starmans is verbonden aan de Faculteit Bètawetenschappen (Department of Information and Computing Sciences) van de Universiteit Utrecht en aan Tilburg University. Hij doet onderzoek op het snijvlak van filosofie, statistiek en informatica.
E-mail: starmans@cs.uu.nl