Counterfactuals Considered Harmful

Jouni Siren

Helsinki 28.3.2003
Kausaalimallien tutkimusseminaari

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

1 Johdanto

Eräs kausaalisen mallintamisen ja päättelyn keskeisimmistä käsitteistä ovat kontrafaktuaalit. Pohjimmiltaan ne ovat väitteitä ”Jos X olisi ollut tosi, niin Y olisi tapahtunut.” tai kysymyksiä ”Mitä olisi tapahtunut, jos X olisi ollut tosi?” Toisin sanoen ne liittyvät spekulointiin, miten tapahtumat olisivat kehittyneet, jos jokin niihin vaikuttava tai sellaiseksi kuviteltu asia olisi ollut toisin. Eräissä toisissa yhteyksissä tällaista kutsutaan jossitteluksi tai jälkiviisaudeksi.

Dawid väittää artikkelissaan [1] kontrafaktuaaleihin perustuvaa kausaalista päättelyä metafyysiseksi ja johtavan herkästi johtopäätöksiin, joita ei voida perustella empiirisin keinoin. Tämä perustuu tosiasiaan, ettei kahden toisensa poissulkevan valinnan seurauksia tietyssä tilanteessa voida kokeilla, vaan ainakin toisen valinnan seuraukset joudutaan arvioimaan aiempien vastaavasta tilanteesta saatujen kokemusten perusteella. Tilalle Dawid tarjoaa päätösteoreettista lähestymistapaa, joka keskittyy eri valintojen hyvyyden arvioimiseen valintatilanteessa käytettävissä olevan tiedon perusteella. Lähestymistavassa on yhteyksiä myös Popperin hypoteettis-deduktiiviseen menetelmään, jossa mallintaja tekee hypoteeseja mallinnettavasta ilmiöstä, johtaa niistä testattavissa olevia seurauksia ja kokeilee, toteutuvatko nämä seuraukset, vai onko malli hypoteesi väärä.

Dawidin keskeinen lähtökohta on, että jos kahta matemaattisesti erillistä mallia ei pystytä empiirisesti erottamaan toisistaan, tulee niistä seurata samat johtopäätökset. Jos johtopäätökset poikkeavat toisistaan, on päättelyä yritetty tehdä mallin väärien, metafyysisten piirteiden perusteella. Tässä tekstissä käsitellään Dawidin näkemysten pohjalta sitä, kuinka kontrafaktuaaleihin perustuva kausaalinen päättely voi johtaa kahdesta empiirisesti ekvivalentista mallista liikkeelle lähdettäessä erilaisiin johtopäätöksiin. Lisäksi katsotaan hieman joitain Dawidin väitteistä esitettyjä kommentteja.

2 Syyt ja seuraukset

Tarkastellaan suurta homogeenista populaatiota U, joka jakautuu erillisiin yksilöihin u. Populaatio U on saatu jakamalla suurempi populaatio U' luokkiin yksilöiden havaittujen relevanttien ominaisuuksien $Z(u)$ mukaan ja valitsemalla jokin näin syntyneistä luokista. Nyt kaikilla yksilöillä $u \in U$ on samat havaitut ominaisuudet $Z(u)$, joten populaation voi katsoa olevan homogeeninen. Muita yksilöiden relevantteja ominaisuuksia ei oleteta olevan tiedossa, vaan päättely perustuu kaikkeen saatavilla olevaan relevantiksi katsottuun tietoon. Populaatiosta valitaan edelleen kaksi riittävän suurta erillistä ryhmää, joista U_t sisältää hoitoa t saavat yksilöt ja U_c on vertailuryhmä. Oletetaan vielä kaiken satunnaisuuden populaation sisällä johtuvan yksilöiden havaitsemattomista piirteistä ja edustavan siten epävarmuutta hoidon vaikutuksesta tiettyyn yksilöön eikä todellista satunnaisuutta.

Artikkelissaan Dawid tarkastelee kahta kausaalisessa päättelyssä keskeistä kysymystyyppiä: mitkä ovat syiden seuraukset (tyyppi I) ja mitkä seurauksien syyt (tyyppi II). Tyypin I kysymys voisi olla esimerkiksi ”Päätäni särkee, auttaisikohan aspiriini?”, kun taas tyypin II kysymys olisi ”Päätäni ei enää särje, johtuukohan se ottamastani aspiriinistä?” Näistä tyypin II kysymykseen vastaaminen edellyttää selvästi kontrafaktuaaleihin turvautumista, sillä niissä kysytään, vaikuttiko tehty päätös havaittuun lopputulokseen, eli olisiko toisenlainen päätös johtanut tässä nimenomaisessa tapauksessa toisenlaiseen lopputulokseen.

Syiden seurauksia kysyvät tyypin I kysymykset ovat sitten toinen asia. Ne eivät sellaisenaan tunnu edellyttävän kontrafaktuaalien käyttöä, sillä kysymyksenasettelu on niissä sama kuin perinteisessä tilastotieteessä: millaisia tuloksia voidaan olemassa olevan tiedon perusteella odottaa, jos päädytään toimimaan jollain tietyllä tavalla. Koska yksilön u saamaa hoitoa t ei ole vielä valittu, ei toiminnan seurauksia kuvaavissa satunnaismuuttujissa $Y_t(u)$ ja $Y_c(u)$ ole mitään metafyysistä. Vasta kun hoito on annettu (tai jätetty antamatta), hukkuu satunnaismuuttuja $Y_c(u)$ (tai $Y_t(u)$) metafysiikan suohon.

Satunnaismuuttujien $Y_t(u)$ ja $Y_c(u)$ jakaumia voi estimoida havaintojen perusteella siinä missä mitä tahansa muitakin jakaumia, jos tehdään sopivat oletukset yksilöiden u samankaltaisuudesta. Sen sijaan parin $(Y_t(u), Y_c(u))$ yhteisjakauman estimoiminen on silkkaa metafysiikkaa: jotta satunnaismuuttujan $Y_t(u)$ arvo voitaisiin havaita, tulee yksilölle u antaa hoito t, joka sulkee pois mahdollisuuden olla antamatta hoitoa ja havainnoida tämän seurauksia. Niinpä kaikki sellainen päättely, jonka johtopäätökset perustuvat oletuksiin parin $(Y_t(u), Y_c(u))$ yhteisjakaumasta, ovat väistämättä metafyysisiä, koska näiden oletusten oikeellisuutta ei voida kokeilla empiirisesti. Käytännössä kuitenkin nähdään usein tyypin I kysymyksiin esitettyjä vastauksia käytettävän tyypin II kysymyksiin vastaamiseen, mikä voi johtaa virheellisiin johtopäätöksiin, jos ei olla tarkkaavaisia.

3 Missä kontrafaktuaalit pettävät?

Kontrafaktuaaleihin yritetään selvittää hoidon t yksilöllinen kausaalinen vaikutus

$$\tau(u) := Y_t(u) - Y_c(u) \tag{1}$$

yksilöön $u \in U \setminus (U_t \cup U_c)$, kun hoito- ja vertailuryhmän tulokset tiedetään. Tarkastellaan nyt tapausta, joka havainnollistaa, kuinka yksilöllistä kausaalista vaikutusta ei voida arvioida minkään havaintoaineiston pohjalta.

Oletetaan satunnaismuuttujien $Y_t(u)$ ja $Y_c(u)$ olevan kaikilla $u \in U$ riippumattomia otoksia normaalijakaumista, joiden odotusarvot ovat θ_t ja θ_c, molempien varianssi ϕ_Y ja korrelaatio ρ. Kun $\rho \geq 0$, saadaan

$$Y_i(u) = \theta_i + \beta(u) + \gamma_i(u), \tag{2}$$

missä $i \in \{t, c\}$ ja $\beta(u)$ ja $\gamma_i(u)$ ovat riippumattomia normaalijakautuneita satunnaismuuttuja, joiden odotusarvo on 0 ja varianssit järjestyksessä $\phi_B := \rho\phi_Y$ ja $\phi_\gamma := (1-\rho)\phi_Y$. Hoidon i vaikutus on siis summa hoidon odotusarvoisesta vaikutuksesta θ_i, puhtaasti yksilöstä riippuvista tekijöistä $\beta(u)$ sekä yksilön ja hoidon yhteisvaikutuksesta $\gamma_i(u)$.

Yhtälöiden (1) ja (2) perusteella yksilöllinen kausaalinen vaikutus voidaan esittää myös muodossa $\tau(u) = \tau + \lambda(u)$, missä $\tau := \theta_t - \theta_c$ on hoidon t keskimääräinen kausaalinen vaikutus ja $\lambda(u)$ on normaalijakautunut odotusarvolla 0 ja varianssilla $2\phi_\gamma$. Havaintojen perusteella parametrit θ_t, θ_c ja ϕ_Y voidaan estimoida luotettavasti. Korrelaation tai yksilöllisen kausaalisen vaikutuksen estimointi ei sen sijaan onnistu, sillä sekä hoidon että hoitamattomuuden vaikutusta samaan yksilöön ei voida havainnoida. Havaintojen perusteella saadaan ainoastaan raja $\phi_\gamma \leq \phi_Y$ (tai $\phi_\gamma \leq 2\phi_Y$, jos ρ voi olla myös negatiivinen). Empiirisesti ei siis voida selvittää, kuinka paljon hoidon tehokkuus vaihtelee yksilöstä toiseen ja kuinka suuri osa vaihtelusta vastaavasti esiintyisi annetuta hoidosta riippumatta.

Merkittävä poikkeus estimoinnin mahdottomuudesta on tapaus $\phi_Y \approx 0$, jolloin hoitotulokset ovat käytännössä riippumattomia yksilöstä. Hoitotulosten deterministisyys selittyy tässä sillä, että populaation U valinnassa käytetyt piirteet Z kattoivat käytännössä kaikki hoidon kannalta relevantit ominaisuudet suuremmassa populaatiossa U'.

4 Päätösteoreettinen lähestymistapa

Kontrafaktuaaleihin perustuvan päättelyn epäonnistuminen äskeisessä esimerkissä ei kerro niinkään siitä, että käytetyt menetelmät olisivat vääriä, vaan että esitetyt kysymykset ovat vääriä. Tyypin II kysymyksen sijasta meidän tulisikin esittää tyypin I kysymys, johon voidaan vastata. Sen sijaan että kysyttäisiin, oliko aspiriinin ottamisesta tällä kertaa hyötyä, tulisi kysyä, oliko päätös aspiriinin ottamisesta silloisten tietojen valossa järkevä tai kannattaisiko aspiriinia ottaa, jos päänsärky uusiutuisi. Jälkiviisastelu ei yleensä ole kovin hyödyllistä, mutta niissä tilanteissa, joissa siihen on turvauduttava, kuten oikeudessa tai onnettomuustutkinnassa, on varottava tekemästä johtopäätöksiä, joita ei edes periaatteessa voida osoittaa vääriksi. Epävarmoissa tilanteissa on tällöinkin parempi turvautua tyypin I kysymykseen ja selvittää syyllisyyden sijasta, oliko tehty valinta tekohetken tietojen perusteella oikea vai ei.

Kun syiden sijasta yritetään arvioida seurauksia, on luontevaa turvautua päätösteoreettiseen lähestymistapaan. Tällöin arvioidaan kaikkien mahdollisten seurausten hyvyys ja valitaan toimintavaihtoehto, joka maksimoi hyödyn odotusarvon tai minimoi häviön odotusarvon. Jos tulkitaan aiemman esimerkin satunnaismuuttujat $Y_i(u)$ päänsäryn kestoksi ja hoito t aspiriinin ottamiseksi on luonnollinen häviöfunktio $L(Y) = Y$. Valinnasta seuraavan häviön odotusarvo on nyt $E[L(Y_t(u))] = E[Y_t(u)] = \theta_t$, jos aspiriinia otetaan, ja $E[L(Y_c(u))] = E[Y_c(u)] = \theta_c$ muuten. Aspiriinin ottaminen siis kannattaa, jos θ_t:n arvon uskotaan havaintojen tai kokeiden perusteella olevan pienempi kuin θ_c. Tämän päätöksen voi katsoa olevan paras mahdollinen, minkä olemassaolevan tiedon (ja funktion L) perusteella voi tehdä.

Vastaava päättely voidaan toistaa kaikissa suuremman populaation U' homogeenisissä osapopulaatioissa U. Huomionarvoista on, että päätöksenteko perustuu pelkästään kunkin yksilön u tunnettuihin ominaisuuksiin $Z(u)$, eikä tuntemattomista mutta mahdollisesti relevanteista ominaisuuksista tarvitse välittää lainkaan. Lisäinformaatiosta ei tietenkään ole haittaa, vaan sellaisen saaminen päin vastoin tarkentaa jakoa osapopulaatioihin ja oletettavasti näin myös arvioita hyödyn tai häviön odotusarvosta. Jos tällaista informaatiota ei kuitenkaan ole tarjolla, sitä ei myöskään tarvitse olettaa.

Päättelyn johtopäätös on sama riippumatta siitä, mikä arvo empiirisin keinoin tavoittamattomissa olevalle parametrille ρ (tai ϕ_γ) on valittu. Sama pätee yleisemminkin: hyötyfunktion odotusarvon pohjalta tehtävät johtopäätökset ovat samat kaikilla niillä malleilla, joita ei voida erottaa toisistaan havaintojen perusteella. Dawid pitää tätä perustavaa laatua olevana vaatimuksena: jos useat mallit vastaavat havaintoja hyvin, eikä valintaa niiden välillä voi tehdä havaintojen perusteella, tulisi lopputulosten kannalta olla yhdentekevää, mihin malliin päätöksentekonsa perustaa.

5 Mitä voidaan päätellä syistä?

Palataan nyt takaisin tilanteeseen, jossa yksilölle u on annettu hoito t ja tämän jälkeen havaittu seuraus $Y_t(u) = y$. Mitä voidaan sanoa siitä, aiheuttiko hoito t seurauksen y tai vaikuttiko se siihen, sortumatta metafyysisten tekijöiden huomioimiseen?

Eräs vaihtoehto on ajatella, että olisi olemassa jokin joukko D ominaisuuksia, jotka määräävät täysin hoidon seuraukset. Tällöin jollain funktiolla f pätee $Y_i(u) = f(i, D(u))$, missä $i = t, c$. Jos vielä pätee, etteivät ominaisuudet muutu hoidon seurauksena, voidaan ne selvittämällä ainakin teoriassa päätellä jotain hoidon yksilöllisestä kausaalisesta vaikutuksesta.

Oletetaan että funktio f tunnetaan ja ominaisuuksille D on pystytty määrittelemään jokin mielekäs priorijakauma. Näiden perusteella voidaan määrittää jakauma $\Pr(Y_c(u) \mid Z(u), Y_t(u))$, joka tarjoaa sitä luotettavamman pohjan päättelylle, mitä suuremmalle osalle Z yksilön ominaisuuksista D tunnetaan arvot. Tyypin II kysymyksiin voidaan siis vastata perustellusti metafysiikkaan turvautumatta, jos tarkasteltavalle ilmiölle tunnetaan deterministinen malli. Jos taas ilmiö ei ole deterministinen tai sen mekanismia ei ole pystytty tunnistamaan, yhtä hyviin tuloksiin ei päästä. Jäljelle jää lähinnä mahdollisuus etsiä uusia selittäviä ominaisuuksia K, joiden avulla yksilöllisen kausaalisen vaikutuksen (1) varianssi ja siten johtopäätöksiin liittyvä epävarmuus pienenee.

6 Muita mielipiteitä

Dawidin artikkeli on ilmestymisensä jälkeen herättänyt runsaasti keskustelua niin puolesta kuin vastaan. Tarkastellaan seuraavaksi joitain keskustelussa esitettyjä näkemyksiä.

Judea Pearl keskittyi omissa kommenteissaan [3] puolustamaan kontrafaktuaalien hyödyllisyyttä päättelyn välineenä. Hän kuitenkin määrittelee kontrafaktuaalit toisin kuin Dawid. Pearlin mukaan tyypin II kysymys ”Olisiko päänsärky kestänyt pidempään, jos en olisi ottanut aspiriinia?” tulisi tulkita kysymyksenä, onko mallista saatu arvio $Y_c(u)$:lle suurempi kuin havaittu arvo $Y_t(u) = y$ [2]. Tämä on kuitenkin tyypin I kysymys eikä siten kontrafaktuaali Dawidin mielessä. Niinpä Pearlin kommentit menevätkin suurelta osin Dawidin väitteiden ohi.

Pearl väittää Dawidin vaatimusta, että matemaattisesti erillisistä malleista pitäisi saada samat johtopäätökset, jos eroa niiden välillä ei pystytä tekemään havaintojen perusteella, merkityksettömäksi tai ristiriitaiseksi. Hän olettaa, että jos johtopäätöksillä ei ole käytännössä näkyviä seurauksia, ei niillä ole merkitystä. Jos taas seurauksia on, voidaan johtopäätösten taustalla olevat oletukset testata seurausten pohjalta, jolloin mallit voidaankin erottaa havaintojen perusteella ja vaatimus on ristiriitainen.

Tämä oletus on kuitenkin virheellinen. Voidaan esimerkiksi kysyä, olisiko jokin lento-onnettomuus tapahtunut, jos lentäjä olisi toiminut toisin. Vastauksella tähän kysymykseen on selvästikin käytännön seurauksia, joten valinta eri tuloksia tuottavien mallien välillä on merkityksellinen. On kuitenkin kyseenalaista, voidaanko johtopäätöksestä seuraavien muutosten lentokoneissa, lentäjien ohjeistuksessa tai jossain muussa asiassa seurauksista päätellä mitään tämän nimenomaisen lento-onnettomuuden aiheuttajasta. Voidaan tietenkin järjestää suuri määrä mahdollisimman samankaltaisia vaaratilanteita ja katsoa, estikö lentäjän toisenlainen toiminta sillä kertaa onnettomuuden. Olosuhteita ei kuitenkaan voida saada täysin samanlaisiksi jo yksin siitä syystä, etteivät alkuperäisen onnettomuuden aikaiset olosuhteet ole täysin tiedossa. Niinpä näistä koeonnettomuuksista saatu tieto ei anna vastausta alkuperäiseen kysymykseen kuin poikkeustapauksissa.

Pearlin kommenttien hyödyllisintä antia on toteamus, ettei kontrafaktuaaleja tulisi kieltää kategorisesti, vaikka niillä saakin helposti virheellisiä tai perustelemattomia johtopäätöksiä. Saattaahan nimittäin olla, että empiirisesti perustelemattomiin oletuksiin perustuva malli osoittautuu myöhemmin oikeaksi, jolloin sen pohjalta tehdyn työn merkitys voi kasvaa huomattavasti. Samoin voi olla, että tällaisesta mallista voidaan tehdä johtopäätöksiä, jotka eivät enää riipu perustelemattomista oletuksista.

Glenn Shafer esittää omissa kommenteissaan [4] enemmän huomautuksia ja tarkennuksia kuin kritiikkiä. Hän keskittyy erityisesti oikeiden ja väärien kysymysten tunnistamiseen. Shafer huomauttaa, ettei eroja havainnoissa saa päätellä yksittäisen valinnan seuraukseksi, jos valintoja tai päätöksiä tehtiin todellisuudessa useita. Jos esimerkiksi hoitoa saavalle ryhmälle ja vertailuryhmälle tehtiin erilaiset lääkärintarkastukset, tulee tämä tarkastus tulkita osaksi hoitoa. Saattaa nimittäin olla, että hoidon seuraukset riippuvat merkittävästi lääkärintarkastuksesta tai tarkastuksen ja hoidon yhteisvaikutuksesta, vaikka tätä ei koetta suunniteltaessa vielä ymmärrettykään.

Oikean kysymyksen valintaan liittyy myös seuraava esimerkki. Henkilö X kuoli 75 vuoden iässä. Hän oli tupakoinut yli 20 vuoden ajan, mutta lopettanut tupakoinnin 40-vuotiaana. On väärin kysyä, kuinka pitkään hän olisi elänyt, jos hän olisi jatkanut tupakointia. Kysymys on niin pitkän aikavälin seurauksista, että moni muukin asia olisi ollut toisin. Selvästi järkevämpi olisi kysyä lopettamispäätöksen vaikutusta hänen odotettavissa olevaan elinikäänsä. Näin ei oleteta, että yksittäinen valinta voitaisiin tehdä toisin kaiken muun pysyessä ennallaan, vaan kysytään selvemmin mitattavissa olevaa asiaa, jota voidaan käsitellä myös Dawidin päätösteoreettisella lähestymistavalla. Lisäksi jälkimmäinen kysymys voidaan tulkita myös tyypin I kysymyksenä, vaikka molemmat kysymykset ovat varsinaisesti tyyppiä II.

7 Lopuksi

Dawid esittää siis, että laajalti kausaalisessa päättelyssä käytetyt kontrafaktuaaleihin perustuvat menetelmät ovat vaarallisia. Niillä saadut johtopäätökset saattavat monissa tilanteissa olla virheellisiä, eikä johtopäätöksiä välttämättä voida varmistaa empiirisesti. Niinpä Dawid vaatiikin, että päättelyn on perustuttava vain mallin niihin ominaisuuksiin, joita voidaan testata empiirisesti. Muiden ominaisuuksien käyttäminen on metafysiikkaa eikä siksi kuulu tieteeseen.

Kontrafaktuaalien tilalle Dawid tarjoaa päätösteoriaan pohjautuvaa ratkaisua, joka tuottaakin hyviä tuloksia monissa tilanteissa. Hänen kategorinen kieltonsa menee toisten mielestä kuitenkin liian pitkälle, sillä kontrafaktuaalit voivat olla hyödyllinen työkalu, vaikka niistä suoraan saadut johtopäätökset olisivatkin kyseenalaisia. Päättelyä ja mallien rakentamista ei kannata rajoittaa pelkästään varmistettaviin menetelmiin, koska toisinaan hyvä arvaus saattaa johtaa parempiin tuloksiin. Toisaalta tällaisten arvauksien pitäisi olla ennemmin poikkeus kuin sääntö, eikä niihin tulisi turvautua ennen kuin konservatiivisempien menetelmien rajat ovat tulleet vastaan.

Lähteet

1 Dawid, A. P., *Causal Inference without Counterfactuals*. Tutkimusraportti, 1997. Myöhempi versio julkaistu kommenttien kera lehdessä *Journal of the American Statistical Association* 95, (2000), sivut 407–448.

2 Galles, D., Pearl, J., An Axiomatic Characterization of Causal Counterfactuals. *Foundations of Science* 3, 1 (1998), sivut 151–182.

3 Pearl, J., The Logic of Counterfactuals in Causal Inference. *Journal of the American Statistical Association* 95, (2000), sivut 428–435.

4 Shafer, G., Comments on ”Causal Inference without Counterfactuals” by A. P. Dawid. *Journal of the American Statistical Association* 95, (2000), sivut 438–442.