Hajautetun tekoälyn ongelmia ja mahdollisuuksia

Raul Hakli

Rationaalisuuden ongelmia

- Havainto 1: Esitettyjen teorioiden (peliteorian, päätösteorian, episteemisen logiikan, bayesiläisen uskomustenpäivityksen, ...) ennustukset käyttäytymisestä eivät päde ihmisten kohdalla. Tämä on todettu useissa testeissä. Ihmiset eivät ole ideaalisen rationaalisia vaan ”erehtyminen on inhimillistä”.
- Havainto 2: Myöskään ohjelmistoagentit eivät ole ideaalisen rationaalisia laskennallisten rajoitusten takia.
- Vastaus: Kyseessä onkin normatiivinen teoria, ei deskriptiivinen.
- Kysymys: Mitä hyötyä siitä sitten on jos haluamme rakentaa järjestelmiä, joissa agentit ovat ohjelmistoagentteja ja ihmisiä?

Peliteorian kritiikkiä

- Väite: peli- ja päätösteoria eivät ole oikea *normatiivinen* teoria rationaalisuudesta. Kuinka väitettä voidaan perustella? Konstruoidaan kuvitteellinen esimerkki jossa agentti rikkoo teorian sääntöjä mutta on kuitenkin meidän käsityksemme mukaan rationaalinen. (Tai vaihtoehtoisesti: konstruoidaan esimerkki, jossa agentti toimii sääntöjen mukaan mutta on silti meidän mielestämme epärationaalinen.)
- Väite: peli- ja päätösteoria eivät ole oikea *deskriptiivinen* teoria rationaalisuudesta. Kuinka väitettä voidaan perustella? Tehdään psykologinen koejärjestely, jossa testataan toimivatko rationaalisiksi oletetut koehenkilöt teorian sääntöjen mukaisesti. Jos ja kun eivät, päätellään että joko teoria on virheellinen tai koehenkilöt (ihmiset) eivät ole rationaalisia.

Normatiivisen teorian testaamisesta

- Kuvitellaan että Janelle tarjotaan mahdollisuutta valita joko 10 € varmasti tai 15000 € mikäli onnistuu heittämään kymmenen kruunaa peräkkäin ja kuvitellaan että Jane ei osaa päättää. → Teorian mukaan Janen preferensseissä nämä tapahtumat ovat yhtä hyvät.
- Muutetaan tilannetta siten että Jane saisikin ensimmäisessä tapauksessa 10 € 5 snt ja oletetaan ettei Jane edelleenkään osaa päättää. Nyt teoriasta seuraa ettei Janella ole preferenssejä senkään suhteen saako hän 10 € vai 10 € 5 snt.
- On kuitenkin luultavaa että hän mieluummin ottaa enemmän rahaa kuin vähemmän. Onko Jane irrationaalinen? (Sugden 1991)

Deskriptiivisen teorian testaamisesta

- Koe (Allais), jossa koehenkilöille annetaan mahdollisuus valita arvonnoissa (A, B) sekä (C, D), missä
  A: [0.8, $ 4000] C: [0.2, $ 4000]
  B: [1, $ 3000] D: [0.25, $ 3000]
- Ihmiset valitsevat B:n mieluummin kuin A:n ja C:n mieluummin kuin D:n.
- Oletetaan että $U(\$0) = 0$. Ensimmäisestä valinnasta seuraa, että $0.8U(\$4000) < U(\$3000)$, mutta jälkimmäisestä päinvastainen epäyhtälö ⇒ ei siis ole olemassa utiliteettifunktiota, joka olisi yhteensopiva näiden valintojen kanssa.
- Havainnolle on vaihtoehtoisia selityksiä. (Tversky & Kahneman 1982).

Deskriptiivisen teorian testaamisesta

- **Ultimatum-peli:** Kahden agentin tehtäväksi annetaan jakaa 100 € kahteen epätyhjään osaan siten että toinen ehdottaa jakoa ja toinen joko hyväksyy tai hylkää jaon. Jos toinen hylkää, kumpikaan ei saa mitään.
- Rationaaliset agentit jakavat rahan siten että ensimmäinen ehdottaa jakoa, jossa itse saa suurimman mahdollisen osuuden (esim. 99 €) ja toinen saa pienimmän mahdollisen osuuden (esim. 1 €). Toinen hyväksyy jaon, koska on parempi saada edes 1 € kuin ei mitään.
- Ihmisille tehdyissä kokeissa näin ei tapahdu vaan epäreiluista jaoista rankaistaan hylkäämällä jako vaikka siitä kärsitään itsekin ⇒ altruistinen rankaisu
- Vaikea tehdä koejärjestelyjä, joissa koetilanne itse ei vaikuttaisi käyttäytymiseen. Joskus se kuitenkin lienee mahdollista, esim. diktaattori-peli tai lompakon kadottaminen.

Mahdollisia selityksiä

Vikaa on etsitty toisaalta ihmisistä ja toisaalta peli ja päätösteorioiden rationaalisuuskäsityksestä

- Ihmiset ovat epärationaalisia
- Teorian oletukset ovat liian vahvoja: ihmisillä ei ole tietoa toisten utiliteeteista eikä rationaalisuudesta
- Ihmisten päätöksentekoon vaikuttavat muutkin kuin rationaalisuuden piiriin kuuluvat seikat (esim. tunteet)
- Ihmisten rationaalisuus on erilaista (esim. kollektiivinen rationaalisuus tai pitkän tähtäyksen rationaalisuus)
- Peliteoria ei huomioi sosiaalisen tilanteen vaikutusta
- ...

Yhteistyön ja altruismin ongelma

Yhteiskuntatieteissä on yritetty selittää miksi ihmiset tekevät yhteistyötä vaikka se ei olisi oman edun mukaista:

- sukulaisten suosiminen (kin selection)
- vastavuoroisuus (reciprocity)
- toistetut pelit (repeated games)
- normit, konventiot, sosiaaliset sanktiot
- maine (reputation)
- empatia ja peilineuronit (mirror neurons)
- ryhmävalinta (group selection)
- kollektiivinen rationaalisuus
- auttamisen ilo
- "aito" altruismi?

Seurauksia tekoälylle

- Agentin utiliteettifunktion tulisi perustua käyttäjän preferensseille, mutta ne eivät ole konsistentteja.
- Rationaalinen agentti, joka toimii peliteorian mukaisesti olettaen että muutkin ovat rationaalisia, ei pärjää (esim. ultimatum-peli tai dominoitujen strategioiden eliminointi)
- Älykäs agentti voi yrittää hyötyä ihmisten "epärationaalisuudesta". Esim. vakuutusyhtiö voi tehdä vakuutustarjouksen, joka ihmisestä vaikuttaa hyvältä, mutta suosii yhtiötä.
- Toisaalta älykäs agentti voi pyrkiä toimimaan ihmismäisesti mahdollistaakseen yhteistyön ja vaikuttaakseen uskottavalta
- Järjestelmän suunnittelijan on otettava kaikki tilanteet huomioon jotta järjestelmä toimisi halutulla tavalla (hyväksikäytön estäminen)
- Joka tapauksessa on tunnettava myös inhimillistä päätöksentekoa eikä pelkkä peli- ja päätösteoria riitä.

Tunnettuja eroavuuksia

Ihmisten preferenssit eivät noudata seuraavia rationaalisuusperiaatteita:

- täydellisyys: $(A \succ B) \vee (B \succ A) \vee (A \sim B)$
- transitiivisuus: $(A \succ B) \wedge (B \succ C) \Rightarrow (A \succ C)$
- kontekstittomuus: preferenssit eivät riipu vaihtoehtojen joukosta

Lisäksi peliteoria abstrahoi pois monia seikkoja, joiden on todettu vaikuttavan ihmisten päätöksentekoon.

- aika
- tieto siitä onko toinen jo tehnyt päätöksen
- vaihtoehtojen esitystapa, "framing effects"
- tieto toisen agentin tyypistä
- "salienssi" jne

Mahdollisia lähestymistapoja

- rajoitettu rationaalisuus (bounded rationality) (H. Simon, Gigerenzer, ...)
- evolutionaarinen peliteoria (Maynard-Smith, Axelrod, Gintis ...)
- behavioraalinen peliteoria (Camerer)
- kollektiivisen intentionaalisuuden teoriat (Gilbert, Searle, Tuomela, Bratman)
- joukkuepeliteoria (team game theory) (Bacharach, Sugden)
- Stackelberg-päättely (Colman ja Bacharach)
- neurotaloustiede (neuroeconomics) (Glimcher & Rustichini)

Rajoitettu rationaalisuus

- Ihmisillä ei yleensä ole aikaa, energiaa tai muita resursseja käydäkseen läpi kaikkia mahdollisia toimintavaihtoehtoja seurauksineen → *maksimointi* on mahdotonta.
- Sen sijaan ihmiset pikemminkin "*satisfisioivat*" (satisfice $\simeq$ satisfy and suffice) ja valitsevat ensimmäisen löytyvän toimintavaihtoehdon joka on riittävän hyvä.
- Ratkaisuvaihtoehtoja ei myöskään käydä läpi systemaattisesti vaan käyttämällä heuristisia "peukalosääntöjä".
- Esimerkiksi yksinkertainen sääntö "win-stay, lose-change" voi johtaa kaikkia hyödyttävään yhteistyöhön ryhmässä, jonka jäsenet eivät edes tunne pelin rakennetta.

Muunnelmia

Myös muita muunnelmia on ehdotettu, kuten Schmidtz (1996): subtiili optimointi (subtle optimization), jossa utiliteettifunktiossa otetaan huomioon laskentaan kuluva aika:

1. Evaluoi seuraava toimintavaihtoehto
2. Arvioi odotettua hyötyä etsinnän jatkamisesta ottamalla huomioon odotetut kustannukset
3. Vertaa parhaan löydetyn vaihtoehdon utiliteetin odotusarvoa etsinnän jatkamisen hyötyyn ja valitse joko paras vaihtoehto tai jatka etsintää

Päätöksenteossa on siis kyse toimintavaihtoehdon etsinnästä ja kuten Tekoäly-kurssista muistetaan, etsintää voidaan tehdä monella tavalla. Perinteistä optimointia voidaan tarkastella sokeana etsintänä ja muita erilaisina heuristisina hakumenetelminä.

Evolutionaarinen peliteoria

- Luovutaan oletuksesta, että pelaajat ovat rationaalisia.
- Sen sijaan ajatellaan että populaatiossa on variaatiota ja erilaisilla yksilöillä on erilaiset genotyypit (vastaavat yksilöiden strategioita).
- Utiliteetit tulkitaan fitness-arvoksi, jolloin luonnon valinta suosii niitä jotka valitsevat "oikein".
- Populaatio päätyy tasapainotilaan, jossa kaikki pelaavat *evolutionaarisesti vakaata strategiaansa* (evolutionarily stable strategy, ESS), joka on paras vastaus sekä itseään että mahdollista mutaatiostrategiaa vastaan.
- ESS-tasapaino on aina myös Nash-tasapaino, mutta kaikki Nash-tasapainot eivät ole evolutionaarisesti vakaita.
- *Replikaattoridynamiikan* (replicator dynamics) avulla voidaan tarkastella populaatioissa tapahtuvia muutoksia.

Kulttuurievoluutio

- Evoluution ei tarvitse olla vain biologista.
- Teorioissa *kulttuurievoluutiosta* (esim. Boyd & Richerson) tarkastellaan uskomussisältöjen, tapojen, rituaalien, normien yms. leviämistä yhteisöissä. Esim. meemiteoriat (Dawkins, Blackmore).
- Oleellisena mekanismina *imitointi*, joka mahdollistaa nopean oppimisen mutta mahdollistaa myös maladaptiivisten piirteiden leviämisen.
- Perimä ja kasvatus ovat vertikaalista transmissiota, mutta ideat ja tavat leviävät nopeasti myös horisontaalisesti. Tärkeimmät imitaation muodot ovat enemmistön matkiminen (conformity-based transmission), menestyneiden matkiminen (prestige-based transmission) ja piirteiden hyvyyteen perustuva matkiminen (content-based transmission).
- Geenit ja kulttuuritekijät vaikuttavat toisiinsa → koevoluutio.

Behavioraalinen peliteoria

- Tavoitteena ihmisten todellisen käyttäytymisen kuvaaminen (siis deskriptiivinen teoria).
- Perustuu empiirisiin havaintoihin (pääasiassa koejärjestelyissä)
- Camerer: "Keskitie ylirationaalisten tasapainoanalyysien ja alirationaalisten adaptiivisten analyysien välillä."
- Teorianmuodostuksen toimintaperiaate: 1) järjestetään koetilanne pelistä, jossa peliteoria tekee selkeän ennustuksen 2) jos havaittu käyttäytyminen poikkeaa odotetusta, kehitä uskottavia selityksiä havaitulle käyttäytymiselle ja 3) laajenna formaalia peliteoriaa näillä selityksillä.
- Kiinnostavaa tekoälyn kannalta, koska toisaalta pyrkii (kuten psykologia) kuvaamaan todellista käyttäytymistä ja toisaalta (kuten taloustiede) tekemään siitä täsmällisiä, matemaattisia teorioita.

Esimerkki

- Koordinaatiopeleissä, joissa yksi vaihtoehto on kaikkien mielestä parempi, peliteoria ei anna selvää ratkaisua, vaikka ihmiset tavallisesti koordinoivat tekonsa parhaaseen ratkaisuun.
- Hi-Lo-peli:

	H	L
H	3,3	0,0
L	0,0	2,2

- "salience"-idea ja "focal points"-teoria (Lewis, Sugden): jotkin vaihtoehdot erottuvat muista.
- *Stackelberg-päättelyssä* agentit olettavat, että toinen vastaa aina parhaalla vastausstrategiallaan ikään kuin osaisivat ennakoida toisen valinnan (Colman, Bacharach).

Kollektiivinen intentionaalisuus

- Filosofisia kysymyksiä: kuinka yksilön intentionaalisuus eroaa ryhmän intentionaalisuudesta, mitä tarkoittaa yhdessä tekeminen, miten se eroaa siitä että moni tekee jotain yksin?
- esim. Searle: juokseminen sadekatoksen alle ja moderni tanssiesitys. Tilanteet eivät eroa ulkoisesti vaan osallistujien mentaalisten tilojen takia.
- Mitä ovat kollektiiviset asenteet (ryhmäuskomukset, yhteiset aikomukset jne) ja onko ryhmillä sellaisia vai ovatko ne vain metaforista kielenkäyttöä?
- useita kilpailevia teorioita siitä kuinka kollektiiviset asenteet tulisi ymmärtää → individualismi vs. kollektivismi (S. Miller, Bratman, Searle, Tuomela, Gilbert)
- Tekoälyssä useita joint intention -ideaan perustuvia formalismeja.

Joukkuepeliteoria

- Joukkuepeliteoriassa tavoitteena on kehittää peliteoriaa kollektiivisen intentionaalisuuden idean pohjalta, jonka mukaan yksilöt tekevät valintoja ryhmän jäsenenä, ei ainoastaan yksilönä.
- Pelaaja ei maksimoi omaa utiliteettiaan, vaan etsii ensin sen lopputuleman, joka maksimoi ryhmän utiliteettia.
- Sitten pelaaja tekee oman osuutensa valitusta tavoitteesta tekemällä sen oman teon, jonka paras lopputulema agentilta vaatii.
- Edellyttää jonkinlaista ryhmäutiliteetin käsitettä, siis kollektiivista intentionaalisuutta. Siitä kiistellään, voidaanko tämä palauttaa yksilöiden preferensseihin.

Yhteenvetoa (1)

- Hajautettu tekoäly, kuten tekoäly yleensäkin, on monitieteinen projekti: useilla tieteenaloilla kehitetään parhaillaan teorioita sekä teoreettisesta että käytännöllisestä päättelystä, päätöksenteosta ja toiminnasta.
- Läheskään aina ei olla tietoisia siitä mitä muilla tieteenaloilla tehdään. Tekoälyssä voi yhdistää monien tieteenalojen teorioita. Joskus vaikutusta myös tekoälystä muiden tieteiden suuntaan.
- Tekoälyn ikuisuusongelmat kuten representaatiotason valinta koskevat myös hajautettua tekoälyä. Toisaalta näyttää siltä että monet ongelmat voidaan kuvata ainoastaan melko vahvalla kielellä, kuten BDI-logiikat, toisaalta kaikenlaista on myös sisäänrakennettuna, kuten peilisoluja (vrt. teoriateoria vs. simulaatioteoria).

Yhteenvetoa (2)

- Ideaalinen rationaalisuus on liian vahva vaatimus ensinnäkin heikon tekoälyn kannalta, toisaalta laskennallisten ongelmien takia ja kolmanneksi epäedullinen valinta järjestelmissä joissa agentit voivat olla ohjelmistoagentteja tai ihmisiä.
- Rationaalisuuden ongelmat vaivaavat lähes kaikkia hajautetun tekoälyn osa-alueita. Toisaalta heuristiset mallit ja muut vaihtoehtoiset lähestymistavat joko kehittymättömiä tai vakiintumattomia. Kaivataan yhdistävää teoriaa.
- Älykkyys (ainakin ihmisen) on luonteeltaan sosiaalista (Tomasello ym.) joten heikon tekoälyn projekti ei voi onnistua olematta hajautettua tekoälyä.
- Sosiaalisuutta ei DAIssakaan ole vielä otettu riittävästi huomioon. Esimerkiksi palauteoppimisen teoriat perustuvat edelleen vain omasta kokemuksesta oppimiseen. Ihminen oppii imitoimalla.

Lähteitä

- C.F. Camerer: Progress in Behavioral game theory, *Journal of Economic Perspectives* 11, 4, 1997, 167–188.
- C. Castelfranchi & R. Conte: Limits of economic and strategic rationality for agents and MA systems. *Robotics and Autonomous Systems* 24, 1998, 127–139.
- A.M. Colman: Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences* 26, 2003, 139–198.
- H. Gintis: Solving the puzzle of prosociality. *Rationality and Society* 15, 2, 2003, 155–187.
- R. Sugden: Rational Choice: A survey of contributions from economics and philosophy. *Economic Journal* 101, 1991, 751-785.
- R. Tuomela: Philosophy and DAI: The Case of Joint Intention, in Jennings, N. and O'Hare, G. (eds.), *Foundations of DAI*, Wiley, 1996.