Kan vi stole på robotene?

Snille roboter er ingen selvfølge. Det kan adferdsøkonomisk forskning lære oss noe om.

Det begynner å sige inn. Robotene tar over. Ikke bare de enkleste rutinejobbene, men også jobbene til DNs lesere. Finansrådgivere og porteføljeforvaltere, advokater og regnskapsførere pustes i nakken av smarte algoritmer og læringsvillige androider. De jobber raskere og de sover ikke.

Men kan vi stole på robotene når interessekonfliktene oppstår? Vil de oss vel?

Vil finansrådgiverroboten gi oss gode investeringsråd, eller vil den anbefale deg å sette penger i dyre fond med tvilsom avkastning?

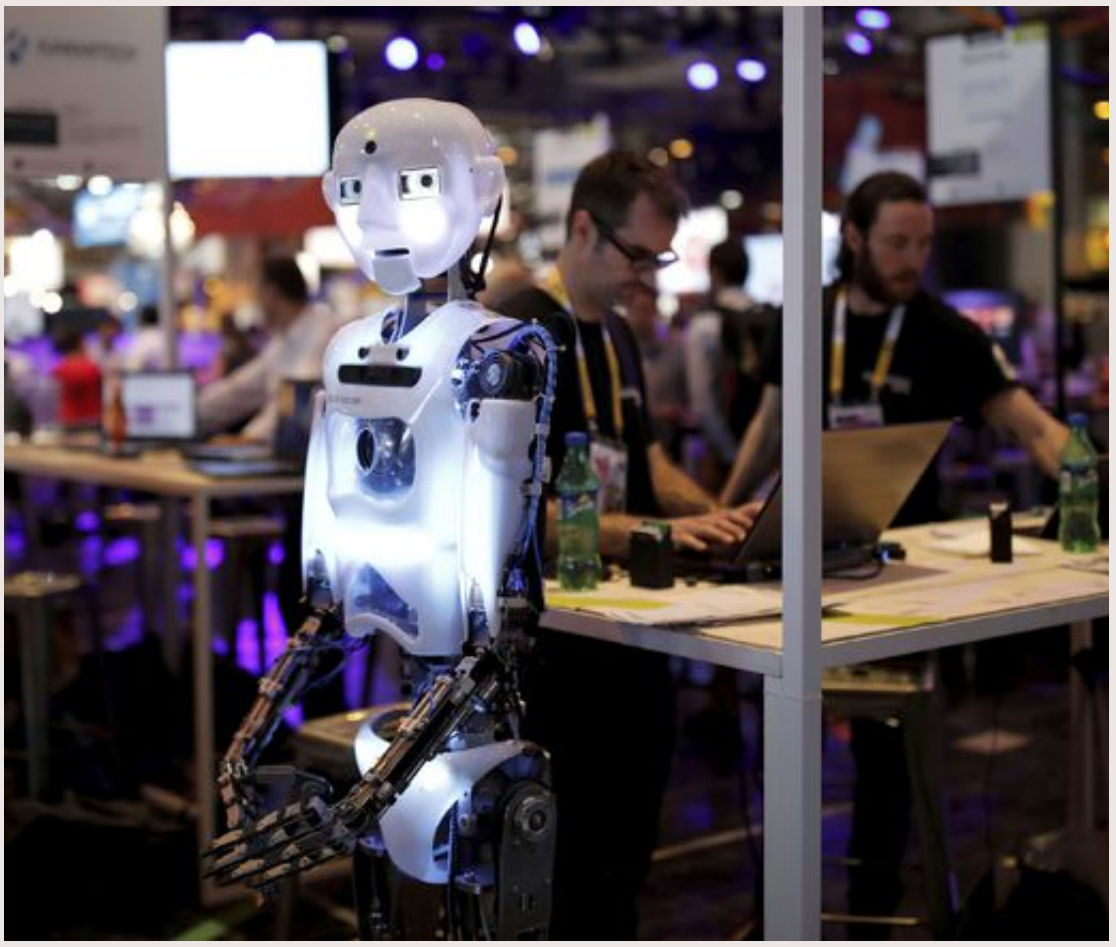
Vil tannlegeroboten ta penger for en unødvendig fylling? Vil revisorroboten holde tilbake viktig informasjon?

Mer generelt: Vil robotene gjengjelde den tilliten du viser dem? Hva er robotenes preferanser?

Mange spørsmål i science fiction sjangeren, dette, men science er det like fullt. Dette er spørsmål som forskere for alvor begynner å jobbe med. Foreløpig domineres feltet av informatikere, men også økonomer og andre samfunnsvitere begynner å fatte interesse for såkalt HRI - «human-robot-interaction».

Økonomene står ikke på bar bakke her. Vi fant opp homo economicus, den rasjonelle følelsesløse aktør, og bør således vite litt om roboter. Men enda viktigere: Adferdsøkonomisk forskning som ikke var ment å studere robotadferd, gir interessante arbeidshypoteser om hva vi kan forvente av robotene.

I adferdsøkonomiske eksperimenter spiller mennesker spill mot hverandre som avslører hvilke «sosiale preferanser» de har, det vil si hvor egoistiske de er, hvor opptatt de er av rettferdighet, eller hvor opptatt de er av å gjengjelde det gode med det gode, og det onde med det onde. Den vanlige metoden er at eksperimentdeltagerne får gjøre sine beslutninger fortløpende etter hvert som spiller skrider frem. Det vil si at de ikke trenger å legge planer på forhånd, men kan i stedet bestemme seg for hva de vil gjøre når de eventuelt får se adferden til dem de spiller mot.



Nyere HRI forskning viser at vi bruker mye av den samme tenkning og strategi mot roboter som mot mennesker. Vi er villige til å gi roboten en sjanse, men vi mister raskt tilliten til den hvis den svikter oss. Foto: Benoit Tessier/Reuters/NTB Scanpix

Men noen ganger pålegges eksperimentdeltagerne å anvende en såkalt strategimetode. Med strategimetoden må spillerne bestemme på forhånd hvilke valg de skal gjøre, gitt de andre spillernes valg.

Ta tillitsspillet som eksempel. En person sender deg et beløp, med det resultat at beløpet tredobles i verdi. Som takk for dette kan du velge å returnere noe til senderen, eller beholde alt selv. Med strategimetoden må du bestemme deg på forhånd for mye du skal returnere til senderen, avhengig av hvor mye du mottar.

Du må altså sette opp en handlingsregel - et valg for hver mulig handling motspilleren velger.

Enn så lenge bruker roboter noe som ligner på strategimetoden. De er programmerte av eierne til å ta beslutninger som en funksjon av alle de ulike situasjonene som kan oppstå. Roboter er riktignok lærende vesener (?) som stadig tilpasser sin adferd til omgivelsene, men forut for enhver interaksjon finnes en form for handlingsregel roboten anvender. Den handler (enn så lenge) ikke på impuls.

Strategimetoden skal i prinsippet ikke gi annen adferd enn den som utfolder seg spontant. Men slik er det ikke. Vi mennesker kan bestemme oss for en ting, men gjøre noe helt annet når situasjonen oppstår.

I tillitsspillet viser det seg at strategimetoden gir betydelige lavere «resiprositet», det vil si at vi returnerer mindre penger til tillitsfulle sendere når vi binder oss til en handlingsregel på forhånd. Vi kan gjerne bestemme oss på forhånd for å være tøffe, men mykner i møtet med tillitsfull eller generøs adferd. Enn så lenge mykner ikke roboten.

Lignende resultater er funnet når man varierer spillernes responstid. Kort betenkningstid har en tendens til å gi mer prososial adferd i en rekke type eksperimenter. Vi deler mer, og bidrar mer til fellesskapet hvis vi ikke tenker oss for lenge om. Den samme intuitive adferden kan vi (enn så lenge) ikke forvente av roboter.

Arbeidshypotesen er med andre ord at robotene i større grad kan være innstilte på å utnytte deg i situasjoner hvor det er en interessekonflikt mellom deg og (eieren av) roboten.

Får vi dermed mindre tillit til robotene? Forskningen så langt tyder ikke på det. I tillitsspillet sender vi mer penger hvis det kun er en algoritme som skal bestemme hvor mye vi får i retur, enn hvis det er et menneske. Dette forklares med «suckeraversion». Vi liker ikke å bli lurt av andre mennesker. Nyere HRI forskning viser imidlertid at vi bruker mye av den samme tenkning og strategi mot roboter som mot mennesker. Vi er villige til å gi roboten en sjanse, men vi mister raskt tilliten til den hvis den svikter oss.

Det er altfor tidlig å konkludere på robotens sosiale preferanser. Denne forskningen er i sin spee begynnelse. Vi vet nok veldig mye mer om noen år, ikke minst når robotene begynner å forske på oss.

Ola Kvaløy, professor i økonomi, Handelshøgskolen ved Universitetet i Stavanger og NHH

Fredagskronikken
Ola Kvaløy

Mer debatt s. 26-28

Økende forskjeller er et av utviklingstrekkene i Norge som bekymrer norske velgere aller mest

Jonas Gahr Støre, leder i Arbeiderpartiet

Debattansvarlig: Vidar Ivarsen **Telefon:** 22 00 10 59 **Sentralbord:** 22 00 10 00 **Epost:** debatt@dn.no **Telefaks:** 22 00 11 10
Hovedinnlegg/kronikk: Maks 4500 tegn inklusive mellomrom **Underinnlegg/replikk:** Maks 1500 tegn (ca. 250 ord) **Legg ved portrettfoto.**
Alt stoff som leveres til Dagens Næringsliv, må produseres i henhold til Vær varsom-plakaten. Dagens Næringsliv betinger seg retten til å lagre og utgi alt stoff i avisen i elektronisk form, også gjennom samarbeidspartnere. Redaksjonen forbeholder seg retten til å forkorte innsendte manuskripter. Debattinnlegg honoreres ikke.