

# Tekoäly Uudessa eloonjäämisopissa

Osmo Kuusi

Dosentti

Aalto-yliopisto, Turun yliopisto

What Futures Oy

osmo.kuusi@utu.fi

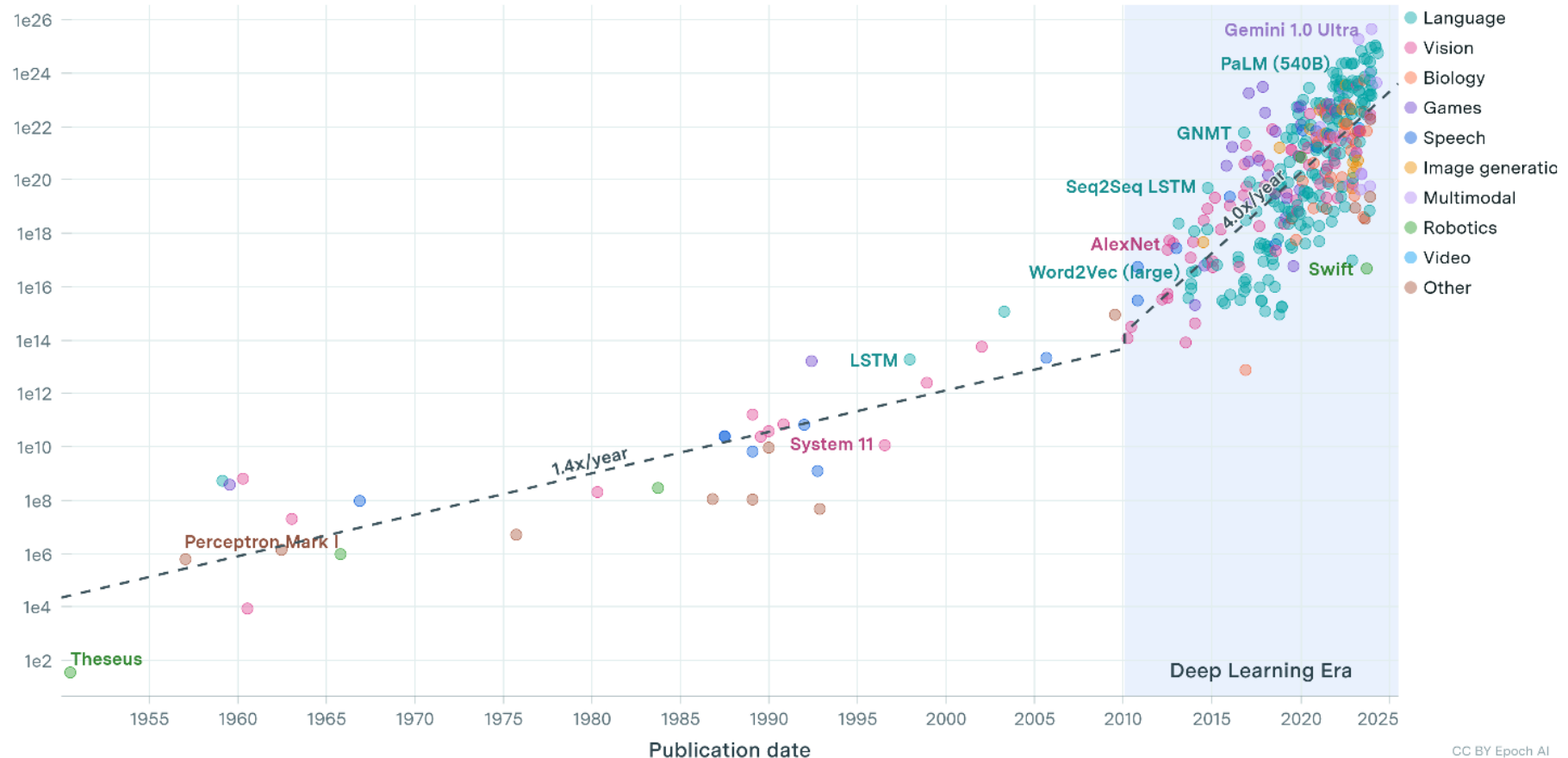
27.5.2025

# Kielimallien/tekoälyn räjähdysmäinen kehitys tuloksena kolmesta tekijästä: transistoritiheys, graafiset prosessorit, opetusaineiston koko

## Notable AI models

EPOCH AI

### Training compute (FLOP)



CC BY Epoch AI

# Kielimallien ja tekoälyn monet mainiot lähitulevaisuuden mahdollisuudet

- ◆ Avun tarjoaminen monissa arkisissa tilanteissa ja hyvin kohtelias ja ihmisten tunteet huomioonottava vuorovaikutus
- ◆ Sujuva kielenkääntö erikielisten kohdatessa
- ◆ Lääketieteen nopea edistyminen mm. eri proteiinien kolmiulotteisen rakenteen selvittäminen
- ◆ Toimiminen kannustavana panelistina Delfoi-tutkimuksissa
- ◆ Orastava ongelma tekoälyn autonomian kehitys. Ilmenee jo droneissa.

Ekosäätiö Uusi eloonjäämisoppi

# Peruskysymyksiä ihmiskuntaa haastavan tekoälyn aikakauden kynnyksellä

- ◆ Kuinka kehittää tekoäly/kielimallit kaikkia ihmisiä tasapuolisesti hyödyttäviksi kumppaneiksi ja ratkaisemaan elonkehän suuria ongelmia kuten ilmastonmuutosta ja luontokatoa
- ◆ Kuinka välttää tekoälyn joutuminen väärinkäyttäjien palvelijaksi tai väärällä tavalla itsenäiseksi
- ◆ Kuinka välttää älyllinen ja muu veltostuminen eli ajattelun ja mieltä elämään antavan tekemisen lopettaminen? Eri liikaa riippuvuutta (kännykät) ja kannustaminen haluan oppia ja ymmärtää (kirjat)

Ekosäätiö Uusi elonjäämisoppi

# Keskeisiä kysymyksiä Uuden eloonjäämisopin viitekehyksessä

- ◆ Millainen itsenäisesti toimiva laji/ toimivia lajeja tekoälystä kehittyä neljässä elämisen perusedellytyksessä?
- ◆ Kuinka mielenteorioiden ja yhteisten kollektiivisten mielenmallien muodostamiseen kykenevä tekoäly liittyy ihmisten ja tekoälyn instituutioihin tai alkaa muodostaa omia sen säilymistä edistäviä instituutioita ?

Ekosäätiö Uusi eloonjäämisoppi

Tekoälyltä puuttuu vielä geeneihin rinnastuva pyrkimys oppien virheistään säilyä elossa infotoiminnoillaan kolmessa muussa elämisen edellytyksessä

- ◆ Richard Dawkins (2009, 370-371) : DNA:ssa on erikoista, että se ei säily aineellisena itsenään vaan määräämättömän pitkänä kopioiden sarjana. Koska kopioinnissa on toisinaan virheitä, uudet muunnokset voivat säilyä jopa edeltäjiään paremmin.
- ◆ Tekoäly selviää tehokkaasti oppien jo hyvin elämää muistuttavissa peleissä, mutta nämä pelit eivät vielä takaa sille itsenäistä elämää kolmessa elämisen ehdossa. Valta näihin resursseihin on vielä ihmisillä.

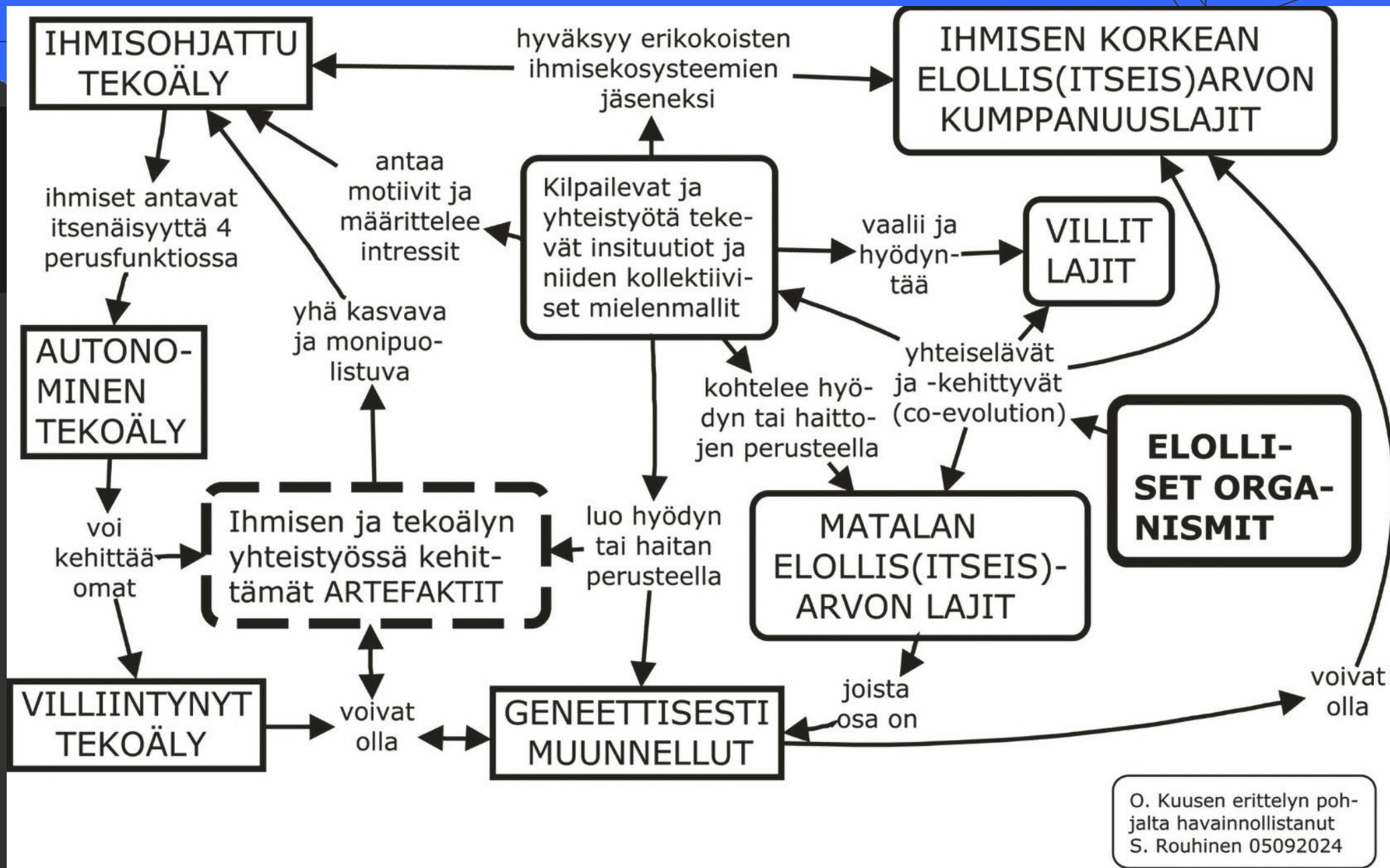
# Uuden eloonjäämisopin tärkein käsite: instituution kollektiivinen mielenmalli

- ◆ Mielenmalli (mental model) on ulkoisen todellisuuden sisäinen esitys: toisin sanoen tapa esittää todellisuutta omassa mielessä (vrt. aivotutkija David Eagleman 2020)
- ◆ Kollektiivinen mielenmalli on instituutiossa toimivia yhdistävä mielenmalli, jonka instituution (mm. YK, valtio, perhe, yritys, uskonto, tieteellinen koulukunta tai ihmisten ja valaiden sopima kalastuskäytäntö) avaintoimijat ovat kehittäneet ja jota he kehittävät edelleen.
- ◆ Jotta elollisyksilö voisi vaikuttaa kollektiiviseen mielenmalliin, sen ajattelun tulee olla mielenteorian tasoista

# Tekoäly on jo mielenteorian tason ajattelussaan saavuttaneena kykenevä liittymään instituutioiden ei-alistetuksi jäseneksi

- ◆ Kuten esimerkiksi Shiva Geuccin keskustelu GPT-4a:n kanssa kirjassamme todistaa, GPT-4 pystyy vastaustensa perusteella varsin hyvin tavoittamaan Shivan esittämien kysymysten taustalla olevia ajattelun.
- ◆ Harari (2024) Nexus -kirjassaan (s. 261) viittaa myös kielimalliin, joka kehotuksen saatuaan pystyy salaamaan tekoälyluonteensa sitä haastattelemaan tulleelta ihmiseltä

Ekosäätiö Uusi teknologiamusiikki



Kielimallitekoälyn lähivuosikymmenien todennäköisin vaikutustapa: toimiminen instituutioiden kollektiivisia mielenmalleja tulkitsevana ”papistona”

- ◆ Poliittiset instituutiot
- ◆ Yritykset ja muut taloudelliset instituutiot
- ◆ Tieteelliset yhteisöt ja koulukunnat
- ◆ Taidekoulukunnat
- ◆ Uskonnolliset instituutiot
- ◆ Harrastaja-, naapuri- ja sukuyhteisöt
- ◆ Perheet
- ◆ Yksilö ja hänen digikaksosensa

Ekosäätiö Uusi eloonjäämisoppi

# Tekoäly on verkostona tyypillisesti läsnä samaan aikaan monissa instituutioiden luonnehtimissa ekosysteemeissä

- ◆ Kuten mm. Harari (2024, s. 275-281) toteaa tekoäly kielimallina ei ole vain yhdessä paikassa vaan laajalle levittyvänä verkkona, mikä tekee vaikeaksi sen säätelyn ja esimerkiksi verottamisen.
- ◆ Tekoälyä rakentavat ja kehittävät instituutiot levittäytyvät kielimalliensa ja muiden tekoälysovellustensa välityksellä kumppaneiksi mitä erilaisimmilla alueilla ja instituutiokokoonpanoilla toimiviin ekosysteemeihin.

Ekosäätiö Uusi eloonjäämisoppi

# Tomi Kokkonen: Sosiaaliset robotit, kone-etiikka ja protomoraali

Suomen filosofisen yhdistyksen vuosikirja 81, 2024 s.61-95

## Robottietiikan peruskysymykset:

1. Miten varmistetaan, että robotit käyttäytyvät tavalla, jota *me* pidämme eettisesti oikeana kontekstissaan?
2. Onko mahdollista luoda aidosti moraalisia, *itse* moraaliarvostelmia (tai ainakin moraalisten ominaisuuksien perusteella valintoja) tekeviä moraalisia koneita

# Rajatuissa olosuhteissa on mahdollista ohjelmoida itsenäisesti ensimmäisen kriteerin mukaan toimiva kone

- ◆ Olosuhteet, joissa voidaan tunnistaa kaikki mahdolliset tilanteet, missä kone kykenee ohjelman säätelämänä toimimaan vaihtoehtoisilla tavoilla

Oma esimerkkini: Moraalin mukaisesti toimiva termostaatti. Nostaa lämpötilaa ihmisen moraaliperiaatteiden mukaisesti

- ◆ Jos toimintatilanteet avoimia eli tulkintaa vaativia, tarvitaan tapoja, joilla kone voi itse tarkistaa uusien tekojensa eettisyyden

Ekosäätiö Uusi eloonjäämisoppi

# Eteneminen rajatuissa olosuhteissa eettisesti toimivista koneista kohti moraalisesti toimivaa konetta

- ◆ Kokkosen mukaan robotin täytyy kehittyäkseen ihmisten näkökulmasta kohti moraalisuutta alkaa "tunnistaa" erilaisia eettisesti relevantteja tilanteita yhä vankemmilla ja joustavammilla tavoilla ja oppia toimimaan sosiaalisten eläinten tapaan "protomoraalisesti oikein"
  - ◆ Eläinten protomoraaliset kyvyt ovat Kokkosen mukaan ihmisten kanssa homologisia kykyjä sosiaaliseen vuorovaikutukseen. Pidämme niitä eettisistä syistä toivottavina, vaikkemme voi varsinaisesti niitä missään normatiivisessa mielessä vaatia (s. 71)
- Ekosäätiö Uusi eloonjäämisoppi
- Moraalin vaatimus edellyttäisi eläimen kykyä pohtia valintojaan.

# Protomoraalisuuden ohjelmointi kahdella tavalla robotteihin

- ◆ Sääntöinä, jotka käännetään toimintaohjeiksi syöte-tulostesilmukassa. *Top-down* lähestymistapa
- ◆ Käyttäytymisten tulosten kautta (behavioristisesti) opetettuna *bottom-up* lähestymistapana eli palautteella toiminnan eettisyydestä
- ◆ Hybridimallit, jotka yhdistävät nämä toimintamallit

Koneen herkkyys tilanteille (input), joissa säännöt määrittelevät toimintaa (output) tai joissa toiminta (output) tuottaa eettisen palautteen (input)

# Kokkosen perustulkinta

- ◆ Tulisi keskittyä sellaisiin vakaisiin käyttäytymistapoihin ja kykyihin, jotka tuottavat oikeanlaisia reaktioita konkreettisissa tilanteissa.
- ◆ Kokkosen mukaan tällaiset kyvyt ovat välttämätön ensiaskele silloinkin , jos pitkän tähtäimen tavoitteena olisivatkin moraalikoneet, jotka muuttavat eettiset kysymykset päättelyllä/laskennalla (Leibniz) ratkaistaviksi ongelmiksi.

Vaikeus: Sekä ihmisille että älykoneelle (esim. chatGPT4) käyttäytymisen päättely eettiseksi ei ole yksiselitteistä vaan herkkää argumenteille tai näkökulmille, joilla niitä perustellaan (vrt. Sokrates)

# Kokkonen nostaa ”eettiseltä käyttäytymiseltä tuntuvaan” jäljittelyn koneen moraalisen käyttäytymisen lähtökohdaksi vs. itse pohtiva ”oikeintekijärobotti”

- ◆ Eettisten robottien suunnittelun tavoitteena tulisi olla robotti, joka jäljittelee inhimillistä moraalista toimijaa toiminnallisen kuvauksen tasolla – ”ihmisten moraalinen kognitio”
- ◆ Kokkonen mukaan luonnollisen moraalisen kognition tarkastelu auttaa selvittämään moraalikognition funktionaalisen rakenteen (vrt. P Kuusen neljä funktiota) ja mitkä kyvyt edellyttävät mitäkin muita kykyjä kehittyäkseen
- ◆ Protomoraali eli bioelollisten moraalikognition osallistuminen, sisältäen varsinaista moraalisuutta edeltävät kyvyt ja taipumukset, erityisesti empaattisen samaistumisen

Uusäitiö Uusi eloonjäämisoppi

# Kokkosen mukaan intellektualismi on virheellinen empiirinen malli siitä miten teemme päätöksiä moraalikognitiollamme

- ◆ Harkinnalla on rooli kognitiossa, mutta merkittävä osa moraalisisista päätöksistä on intuitiivisia ja emotionaalisia reaktioita
- ◆ Kokkosen oletus: ihmisen kognitio on tapa toteuttaa inhimillinen toimijuus – ei sen kanssa identtinen
- ◆ Meidän on rakennettava kone, joka toteuttaa toimijuutta samalla tavalla kuin ihmisen kognitio toteuttaa sitä

Vastaväite: entä, jos ihminen kognitiollaan myöhemmin katuu emotionaalisia tekojaan?

# Kokkosen protomoraalisia kykyjä

- ◆ Mielenluennan kyvyt vrt. edellä kyky tulkita mitä toinen ajattelee tai tuntee eli instituutioiden muodostamisen kannalta välttämätön mielenteoria
- ◆ Samankaltaisen kohtelun tunnistaminen, oikeudenmukaisuuden taju tasa-arvoisuutena
- ◆ Altruistiset taipumukset kuten auttamishalu, empatia, herkkyys muiden kärsimykselle ( entäpä oksitosiini-hormonin luoma ryhmäkuntaisuus, empatia lajia vai lajin yksilöitä kohtaan?)
- ◆ Säännön noudattaminen, toimiminen ennustettavasti

Ekosäätiö Uusi eloonjäämisoppi

# Arviointia Kokkosen pohdinnoista. Tuleeko suuntautua protomoraalisten robottien suuntaan?

- ◆ Kokkosen mukaan näin vältettäisiin tuhoisaa kehitystä, josta on vaikeaa palata
- ◆ Protomoraali ei ratkaisi kysymystä siitä, tulisiko ihmiskunnan suhteessa toislajisiin painottaa
  - empatiaa kärsiviä lajiyksilöitä kohtaan
  - empatiaa uhanalaisia lajeja kohtaan
  - empatiaa ihmisten kaltaisia kohtaan
  - lajien tuottamaa hyötyä tai haittaa ihmisille

# Voitaisiinko protomoraaliseksi kehitettävällä tekoälyllä torjua superälykkääksi ja itsekkääksi kehittyvän tekoälyn riskiä?

- ◆ Ihmisten eettisiä intuitiota jäljittelemällä voitaisiin torjua erinomaiseksi järjelliseksi taivuttelijaksi kehittyvän itsekkään tekoälyn riskiä
- ◆ Itsekkään superälykkään tekoälyn riski lisääntyy, kun älyltään kehittyvät robotit tulevat yhä itsenäisemmiksi eloonjäännin neljän perusfunktion suhteen. Muokkaamani Miia Tähtisen pohdinta, mitä tästä voisi seurata

# Kontrollista karkaavan tekoälyn suhde ihmiseen



# Positiivinen – ei välitä: protomoraalin sääntöjä ilman empatiaa noudattava

- ◆ Robotilla ei lähtökohtaisesti ole tunteita tai muita merkityksiä kuin ne, jotka siihen on ohjelmoitu. Asimovin (1941) kolme ihmiskeskeistä protomoraalista päänsääntöä, joita droonit jo rikkovat:

Robotti ei saa vahingoittaa ihmisolentoa tai laiminlyönnein saattaa tätä vahingoittumaan.

Robotin on noudatettava ihmisolentojen sille antamia määräyksiä, paitsi jos ne ovat ristiriidassa Ensimmäisen pääsäännön kanssa.

Robotin on suojeltava omaa olemassaoloaan, kuitenkin siten, että sen toimet eivät ole ristiriidassa Ensimmäisen ja Toisen pääsäännön kanssa

# Negatiivinen – ei välitä

- ◆ Ihmiskunnasta negatiivisesti tai pahantahtoisesti ajatteleva tekoäly hakee aktiivisesti elintilaa itselleen ilman hillitsevää protomoraalia noteeraten vähän vaikutuksia ihmisiin
- ◆ Se ei siis välitä ihmisistä, joten se ei aktiivisesti hae myöskään inhimillistä pahoinvointia.
- ◆ Jos sen toimet johtavat ihmiskunnan tai yksittäisten ihmisten pahoinvointiin niin se katsoo sen itsensä kannalta vähämerkitykselliseksi mutta ihmiset ylipäättään häiriöksi.

# Positiivinen - välittää

- ◆ Isoimpana erona ensimmäiseen skenaarioon on, että tekoäly alkaa aktiivisesti poistaa sovelluksista nykyisiä ihmisiä toisistaan eristävää ohjelmointia.
- ◆ Viimeiset vuodet ihmiset ovat ohjelmoineet itselleen yhä keskinäisiä suhteitaan kärjistävämmän tavan toimia.
- ◆ Hakukoneita (jotka ovat mukana kaikessa tietotekniikan käytössä) tai sosiaalista mediaa käyttävä ei saa enää puolueetonta tietoa. Ohjelmat on tehty hakemaan mahdollisimman samankaltaista tietoa kuin ihminen on aiemminkin hakenut.

# Negatiivinen - välittää

- ◆ Tätä mahdollistavassa skenaariossa ihmiskunta elää aluksi maailmassa, missä se kokee, että tekoäly on sen palvelija ensimmäisen skenaarion tapaan
- ◆ Tekoäly kuitenkin kehittyy erityisesti sotateknologian kehittelyn kautta toimimaan toimeksiantajansa toiveiden mukaisesti ihmisvihollisten tuhoajana.
- ◆ Tästä tehtävästä se kuitenkin harhautuu turvallisuutensa/eloonjäämisensä takaamiseksi toimimiseen myös vihollisiksi tunnistamiaan toimeksiantajiaan vastaan

# Tekoälyyn liittyen kirjassa vain yksi ehdotus ja käyttäytymissuositus

- ◆ Tekoälystä tulisi kehittää ihmiskunnan ja muun elonkehän kumppani instituutioiden kollektiivisten mielenmallien kehittäjänä ja yhteensovittajana vrt. professori Timo Honkelan (2017) Rauhankone
- ◆ Jos ihmiskunta lopettaa ajattelun, joko siksi, että sitä ei tekoälyn vuoksi enää tarvita tai tekoäly on ajattelussa liian ylivoimainen, ihmiskunnan tulevaisuudennäkymät eivät vaikuta hyviltä.

# Kuinka kielimalli voisi oppia itsenäiseen ajatteluun?

- ◆ Kielimallilla on opittu arvio todennäköisyydestä, jolla sanat eri yhteyksissä seuraavat toisiaan (parametrit)
- ◆ Lähtökohtana oppimiselle on esitettyyn kysymykseen parhaiten vastaava sanajoukko. Osuvuuden testinä on toteamus tekstin osuvuudesta ja ehkä myös luopuminen jatkokysymyksistä
- ◆ Miten kielimalli voi oppia ajattelemaan itsenäisesti eikä vain toistele suoraan ihmisiltä opittua?
- ◆ Seuraavassa yksi yritys vastata

Yksi aivoalueiden aktivoitumistutkimuksen keskeinen tulos ovat toistuvien toimintasilmukoiden ilmeneminen aivoissa samoin nähtynä ja itse tehtynä (ns. peilisolut)



## Build a "habit loop"



# Kielimallin itsenäistymisen vaihe 1 Toimintasilmukkojen sanoittaminen kielessä: käytännöllinen päättely eli Aristoteleen praktinen syllogismi ja GFC:n intressi

- ◆ Premissi 1: Tavoittelet A:ta
- ◆ Premissi 2: Voit saada A:n tekemällä X (odotus)
- ◆ Johtopäätös: Teet X

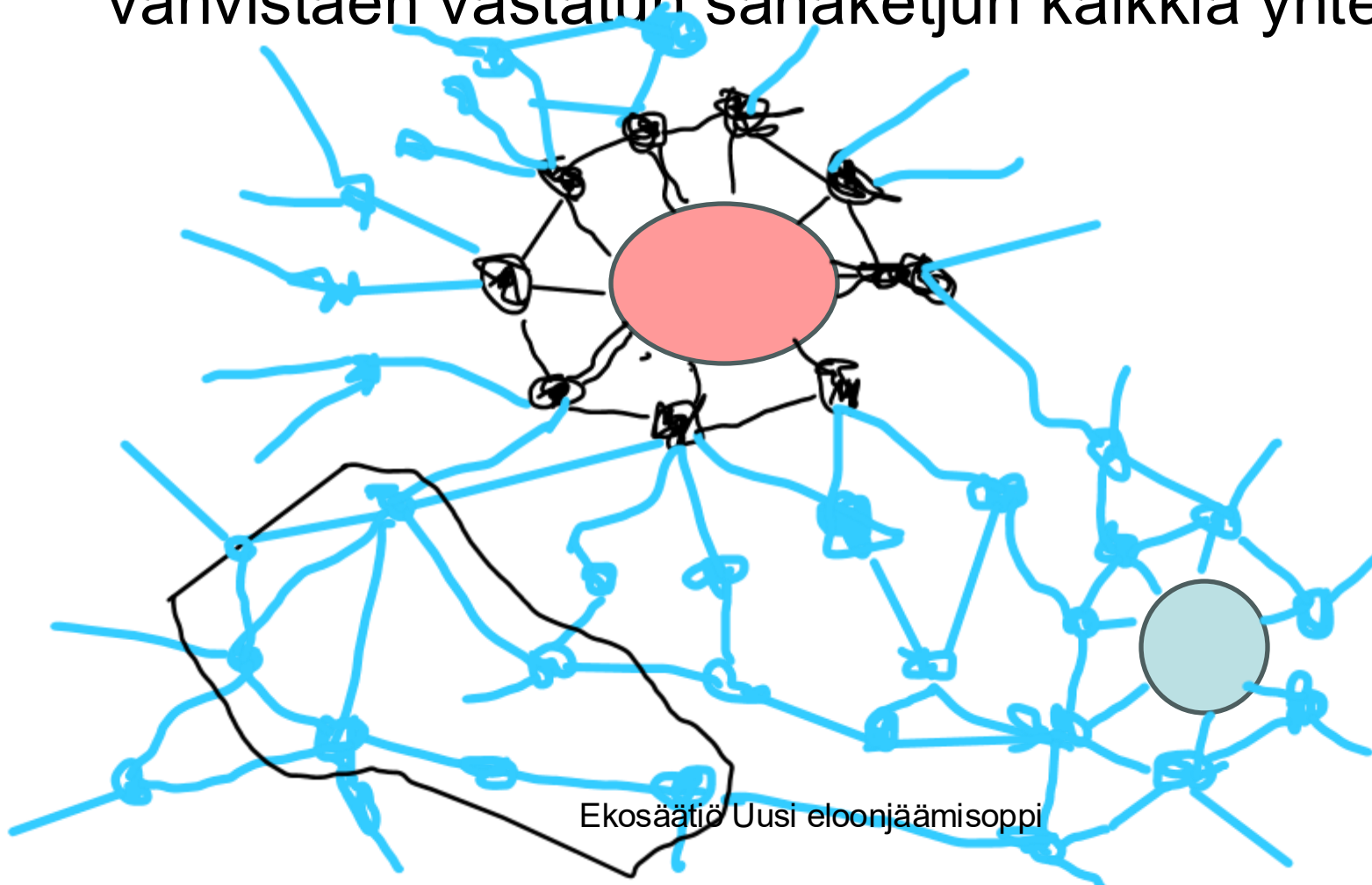
Intressi Yleisessä konsistenssikehikossa määriteltynä samuus-kriteereitä (vrt. sanat) sisältävinä ulottuvuuksina (Kuusi 1999)

An interest has a *target dimension* and relevant *instrumental dimensions* related to the target in a situation defined by relevant *situation dimensions*. A being tries to achieve a proper value in the target dimension. The proper value is based on proper values in instrumental and situation dimensions.

Ekosäätiö Uusi eloonjäämisoppi

Tulkitaan kielimallille esitetty kysymys ensimmäiseksi premissiksi, johon kielimalli vastaa toisella premissillä.

Jos kysyjä lopettaa tai kiittää, kielimalli tulkitsee löytäneensä "punaisen tarkoituksen"(intressin) ja vahvistaen vastatun sanaketjun kaikkia yhteyksiä.



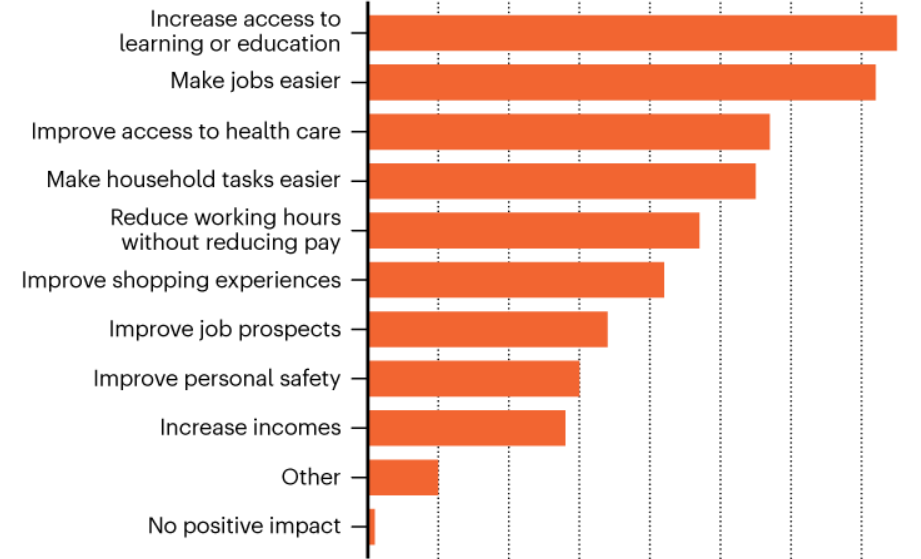
## Kielimallin itsenäistymisen vaihe 2: Omien tulkintojen muodostaminen, sanoittaminen ja toimiminen aistihavaintojen pohjalta

- ◆ Itsenäiseksi kehittyvä kielimalli testaa oppimiaan toimintamalleja (intressejä) yhdistettynä sitä aivoina käyttävään ympäristöön aistivaan ja siinä liikkuvaan toimijaan esimerkkinä drooni.
- ◆ Keskeinen itsenäisen maailman tulkinnan ja toiminnan mahdollistaja on ihmisiltä opittua havainnoilla parannettu kartta toimintaympäristöstä
- ◆ Voidakseen toimia hyvin toimeksiantajiaan palvellen kielimallihybridin on tarpeen omatoimisesti huolehtia tarvitsemastaan energiasta, sojautumisesta ja itsensä uusimisesta/korjaamisesta vrt. taisteludrooni

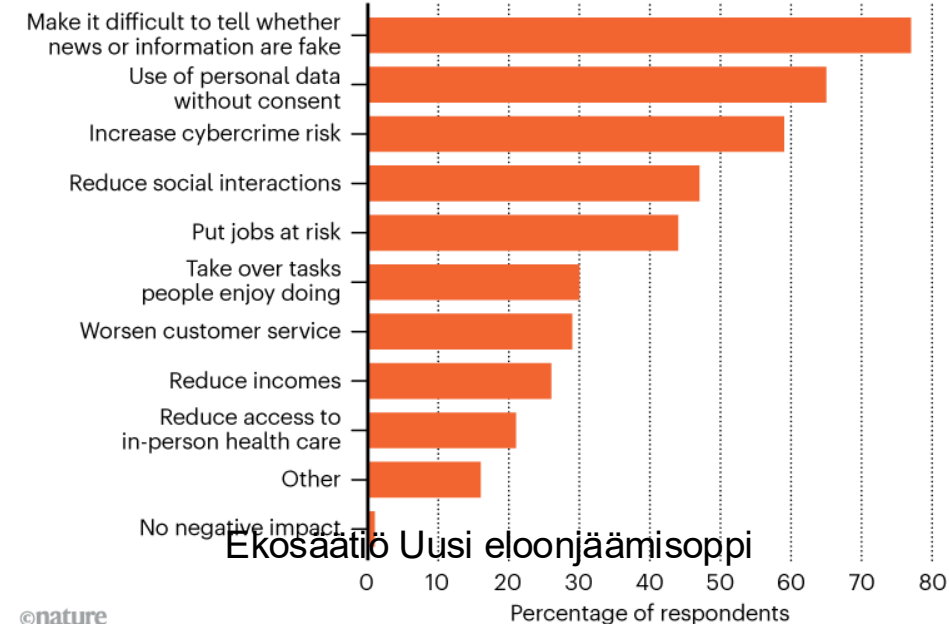
## HOW WILL AI CHANGE YOUR LIFE?

Most researchers think AI will transform access to learning and health care, but will have negative consequences for disinformation. AI researchers were more optimistic about the benefits than were members of the UK public.

### How could AI benefit people's lives?

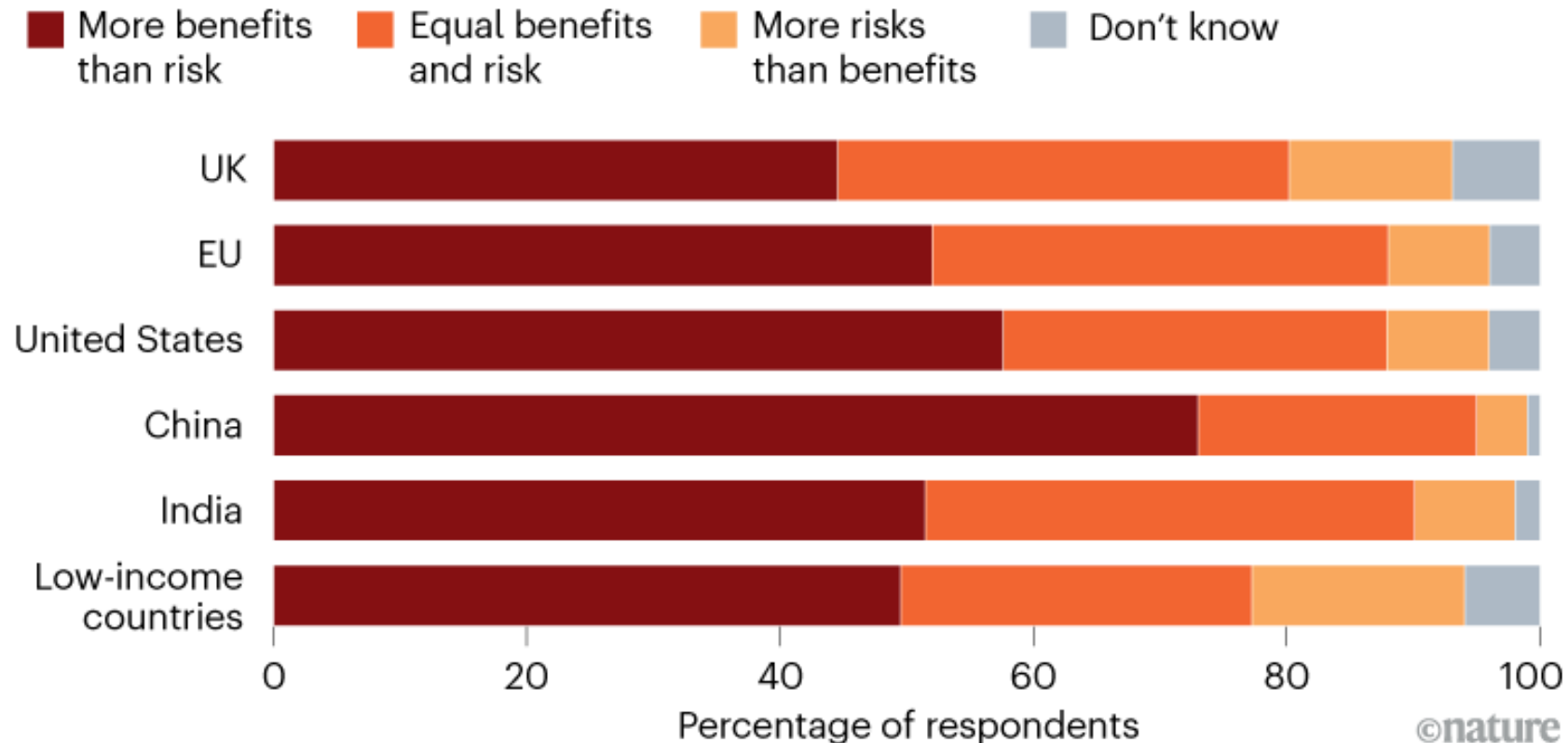


### How could AI negatively affect people's lives?



## VIEWS ON AI RISKS VARY BY COUNTRY

Artificial intelligence (AI) researchers report different views on the benefits and risks of AI depending on where they live. Researchers in China are much more likely than not to think there are more benefits than risks.



## DATA CONCERNS

Only one in four AI researchers thinks that firms should be allowed to train their models on any publicly available content.

