1. Tekoälyn perusteista [RN95, osa I][Nil98, luku 1]

Tekoäly on *uusi* tieteenala: nimensä se sai 1956. Se syntyi vain noin 10 vuotta ensimmäisten yleiskäyttöisten tietokoneiden jälkeen, joten se on yksi vanhimpia alueita *tietojenkäsittelytieteen sisällä*.

Se on samalla *vanha* tieteenala: yli 2000 vuotta on pohdittu, miten havaitsemista/oppimista/muistamista/päättelyä/. . . tehdään/tulisi tehdä.

Toisin kuin vanhemmat sisarensa *filosofia* ja *psykologia*, tekoäly pyrkii *rakentamaan* älykkäitä olentoja/yksikköjä/koneita/. . . eikä (pelkästään) ymmärtämään itse älykkyyttä.

Tämän luvun eri asiat liittyvät monin eri tavoin toisiinsa, joten se pakostikin sisältää runsaasti ristiviittauksia.

1.1. Käsitteen "tekoäly" merkityksiä

Jotta tekoälystä voi "filosofoida", pitää ensin määritellä, mitä sanalla tarkoittaa.

Eri määritelmät johtavat erilaisiin tutkimustavoitteisiin ja -menetelmiin.

Eräs mahdollinen jako saadaan tekemällä yhtä aikaa kaksi erottelua: *ajatella/toimia* ja *inhimillisesti/rationaalisesti*. Näin syntyy nelikenttä:

"The exciting new effort to make computers think ... *machines with minds*, in the full and literal sense" (Haugeland, 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning ..." (Bellman, 1978)	"The study of mental faculties through the use of computational models" (Charniak and McDermott, 1985) "The study of the computations that make it possible to perceive, reason, and act" (Winston, 1992)
"The art of creating machines that perform functions that require intelligence when performed by people" (Kurzweil, 1990) "The study of how to make computers do things at which, at the moment, people are better" (Rich and Knight, 1991)	"A field of study that seeks to explain and emulate intelligent behavior in terms of computational processes" (Schalkoff, 1990) "The branch of computer science that is concerned with the automation of intelligent behavior" (Luger and Stubblefield, 1993)

Taulukko 1: 4 käsitystä tekoälystä [RN95, kuva 1.1].

Eri lokeroiden väliset raja-aidat eivät ole tiukkoja:

- Kummankin toiminnan painottaja edellyttää (tietenkin) myös kykyä vastaavaan ajatteluun.
- Ajattelun motivaationa on kääntäen tarve myöhemmin toimia.
- Ainakin toivoisi, että inhimillisyys ja rationaalisuus olisivat edes kaukaista sukua toisilleen. . .

Lokeroinnissa onkin kysymys siitä, mitä pitää *itse tavoitteen saavuttamisena* ja mitä sitä edeltävänä/seuraavana lisätyönä.

1.1.1. "tekoäly" = "toimia inhimillisesti"

Taulukon 1 vasemman yläkulman tunnetuin muotoilu on *Alan Turingin (1912-1954)* vuonna 1950 esittämä "oikeussalidraama", jossa konetta syytetään "väärennetyillä henkilöllisyyspapereilla esiintymisestä":

syytetyistä $\mathbf{A}$ ja $\mathbf{B}$ toinen on ihminen ja toinen ihmistä teeskentelevä kone M.

ihmistuomarin $\mathbf{C}$ tehtävänä on päättää kumpi on kumpi. Hän esittää valitsemiaan kysymyksiä syytetyille ja vastausten perusteella julistaa tuomionsa.

"Oikeus on sokea": tuomarilla C ei ole muuta informaatiota (tuntoa, näköä, hajua,. . .) syytetyistä kuin kysymysten kirjalliset vastaukset.

Konetta on pidettävä älykkäänä, jos C erehtyy tuomiossaan.

Tai tilastollisesti: tuodaan sama kone "rikoksenuusijana" yhä uudelleen saman C eteen eri ihmisten kanssa, ja kutsutaan sitä älykkääksi, mikäli yhteenlaskettu virheiden määrä on samaa luokkaa kuin jos C olisi kuulustelujen sijaan heittänyt kruunaa ja klaavaa.

periaate: kone on "älykäs" jos sen *syöte-tuloste -käyttäytymistä* ei voida erottaa luotettavasti ihmisen vastaavasta.

Turing ottaa näin *funktionalistisen* kannan älykkyyteen. Myöhemmät tämän kannan edustajat ovat usein siirtyneet tarkastelemaan kokonaisuuden sijasta sen *osien* käyttäytymistä esimerkiksi väittämällä, että vain yksittäisten neuronien (hermosolujen) syöte-tulossignaalit merkitsevät, ei niiden välittämisessä käytetyt biokemialliset mekanismit.

Voisiko funktionalismin hylätä rajaamatta "älykkyyttä" samalla vain tietynlaisiin eliötyyppeihin (joita saattaa esiintyä vain maapallolla)?

1.1.2. "tekoäly" = "ajatella inhimillisesti"

Ihmisen mielen
miten kuvaisit? Se on
tushipiirroksen
mäntyjen oksistossa
tuulevan tuulen ääni.

Ikkyū (suom. Tuomas Anhava)

Rajusti kärjistäen taulukon 1 vasenta alakulmaa, *kognitiotieteen (Cognitive Science)* menetelmä älykkyyden mallintamiseen on seuraava:

1. Kerätään ihmisryhmältä I *kokeellisen psykologian* menetelmin havaintoaineistot H ja H' sen jäsenten käyttäytymisestä jossakin "älykkyyttä vaativassa" tehtävässä T.
2. Rakennetaan sellainen *laskennallinen malli (ohjelma)* M, joka antaa (likimain) samat tulokset kuin lähde- tai opetusaineisto H.
3. Testataan, antaako M myös (likimain) samat tulokset kuin testiaineisto H'.
4. Jos antaa, niin tämä tukee seuraavia hypoteeseja:
   (a) M ratkaisee tehtävän T "älykkäästi", käyttäytyyhän se samaan tapaan kuin joukon I jäsen; ja kääntäen
   (b) M:n laskennalliset operaatiot muistuttavat niitä mentaalisia operaatioita, joita joukon I jäsen käyttää ratkoessaan tehtävää T.

periaate: kone on "älykäs" jos siinä tapahtuu *samankaltaista informaationkäsittelyä* kuin ihmismielessä.

Luvun 1.1.1 Turingin testi -määritelmään verrattuna ei "onnistumisen" kriteeri ole enää yhtä selvä, mutta toisaalta nyt päästään vetoamaan mallin/koneen M *sisäiseen* rakenteeseen ja toimintaan.

1.1.3. "tekoäly" = "ajatella rationaalisesti"

logic *n.* The art of thinking and reasoning in strict accordance with the limitations and incapacities of the human misunderstanding.

A. Bierce

Filosofiassa ja matematiikassa on tutkittu paljon *logiikkaa*, "oppia oikeasta ajattelusta": sääntöjä, joita järjestelmällisesti seuraamalla ajatteluun ei pääse putkahtamaan virheitä.

Esimerkiksi *Aristoteleen (384–322 eKr.) syllogismit* kuten

$$\text{Sokrates on ihminen.} \tag{1}$$

$$\text{Kaikki ihmiset ovat kuolevaisia.} \tag{2}$$

$$\text{Siispä Sokrateskin on kuolevainen.} \tag{3}$$

tuntuvat "järkeviltä": jos olen jo vakuuttunut väitteistä (1) ja (2), niin johdonmukaisuuden nimissä minun täytyy silloin hyväksyä myös johtopäätös (3). Minun täytyy jopa hyväksyä kaikki *saman muotoiset* perustelut, joissa 'Sokrates', 'ihminen' ja 'kuolevainen' korvataan muilla sanoilla.

periaate: kone M on "älykäs" jos se maailman tilan ja säännönmukaisuuksien kuvauksesta pystyy *päättelemään* onko annettu tavoite saavutettavissa (ja jos, niin miten).

Toisin kuin lukujen 1.1.1 ja 1.1.2 määritelmissä, taulukon 1 oikeassa yläkulmassa pyritään ihmisen jäljittelyn sijasta "oikeaan lopputulokseen". Tämä kuitenkin määritellään koneen M *ulkopuolisin kriteerein* U, kun taas aiemmin tavoite "olla kuin ihminen" oli "sisäistetty".

Kun tavoite on ulkoa annettu, voi "tietämysinsinööri" *(knowledge engineer)* alkaa *suunnittelemaan ja rakentamaan* sen toteuttavaa konetta, jota sitten voisi väittää älykkääksi.

1.1.4. "tekoäly" = "toimia rationaalisesti"

Kun vaaditaan vielä luvun 1.1.3 koneelta M *toimintaa* muodostamiensa rationaalisten ajatusten mukaisesti, saadaan oppikirjojemme [Nil98, RN95] kanta, jonka mukaan tekoäly on *älykkäiden toimijoiden (intelligent agents)* suunnittelemista ja rakentamista.

Toimija (agent) on mikä tahansa erillinen yksikkö A, jonka voidaan ajatella tekevän *havaintoja (percept)* ympäröivän maailman *tilasta (state)* sekä *teoillaan (action)* vaikuttavan siihen.

Informaation kerääminen havainnoilla on sekin tekemistä!

Suorituskykymitta (performance measure) U on luvussa 1.1.3 mainittu toimijan A ulkopuolinen mittari sen (pitkän aikavälin) toiminnan tuloksellisuudelle. Ajatellaan se (numeerisena) funktiona *ympäröivän maailman eri tiloille* S: siivousrobotilla $U(S) < U(S')$, jos S' on puhtaampi maailma kuin S.

Esimerkillinen rationaalinen toimija (ideal rational agent) A_{ert} valitsee jokaisessa tilassa S_{nyt} sellaisen mahdollisen teon T_{nyt}, joka maksimoi suorituskykymitan U *odotusarvon (expected value)*.

Odotusarvon, koska (esimerkiksi autoilevan) toimijan A_{ert}

- lisäksi voi olla *muita* toimijoita
  "Tuleekohan tuolta sivutieltä ketään eteeni?"
- *tieto* nykytilasta S_{nyt} tai tekojen T_{nyt} vaikutuksista voi olla epävarmaa
  "Onkohan tie liukas?"
  "Kiihtyyköhän autoni tarpeeksi ohitukseen?"
- tekojen T_{nyt} *onnistuminen* voi olla epävarmaa
  "Entäs jos pyörät sutivatkin tyhjää kun kaasutan?"

Liikennesääntö "risteysmerkin jälkeen ei saa enää ohittaa" voidaankin nähdä laskutoimituksen *yksinkertaistuksena* karsimalla vaihtoehtoja.

Valinnan T_{nyt} pohjana A_{ert} käyttää

- sisäänrakennettua tietoa

  Liikennesäännöt, . . .

- ennen nykytilaa S_{nyt} keräämäänsä *havaintojonoa (percept sequence)* H_{nyt}.

  Onko auto liukastellut tänään aiemmin? Miten auto on ennen kiihtynyt? . . .

Jonon H_{nyt} perusteella A_{ert} muodostaa oman käsityksensä siitä, millainen S_{nyt} on.

A_{ert} on *itseohjautuva (autonomous)* siltä osin, kuin sen käyttäytymisen on sen aiempien kokemusten H_{nyt} määräämää.

Kellokin on esimerkillinen rationaalinen toimija – jos tavoitteena on vain pyörittää viisareita tahdissa antamatta minkään häiritä niiden kulkua. Se ei havainnoi ympäristöään eikä siis ole lainkaan itseohjautuva, vaan sen "rationaalisuus" on kokonaan kellosepän siihen istuttamaa.

A_{ert} ei saa ummistaa silmiään *olennaiselta* informaatiolta: ennen ohitusta kannattaa katsoa taustapeiliin, ettei käänny takaa tulevan ohittajan eteen – maantieajossa ohituskolarin U-arvo ei liene järin suuri. . .

. . . mutta vuoden 1997 viimeisessä Formula 1 -osakilpailussa Schumacher *toimi järkevästi* kiilatessan Villeneuvea vastaavassa tilanteessa:

Villeneuve pisteittä: Schumacher on mestari.

Villeneuve ohittaa: Schumacher ei ole mestari.

- $U_1 = U_{\text{Schumi}}(\text{Schumacher päästää Villeneuven ohitseen}) \ll 0.$
- $U_2 = U_{\text{Schumi}}(\text{Schumacher kiilaa Villeneuven(kin) ulos}) \gg 0.$
- $U_3 = U_{\text{Schumi}}(\text{Schumacher kiilaa Villeneuven taakseen}) > 0.$

- $U_4 = U_{\mathsf{Schumi}}(\text{Schumacher kiilaa vain itsensä ulos}) \ll 0$.
- Todennäköisyys $P_1 = P(\text{Schumacher kiilaa Villeneuven ulos})$ on hyvin suuri, onhan hän asiansa osaava kuski. . .
- Samoin $P_2 = P(\text{Schumacher kiilaa itsensä ulos})$ on hyvin pieni.

Silloin kiilaamisen odotusarvo on

$$P_1 \cdot U_2 + (1 - P_1) \cdot (1 - P_2) \cdot U_3 + \underbrace{(1 - P_1) \cdot P_2}_{\text{hyvin, hyvin pieni}} \cdot U_4$$

mikä lienee (paljon) suurempi kuin U_1.

Sillä kertaa kuitenkin tuo pessimaalinen vaihtoehto toteutui.

Vaikka jokaisella A_{ert} onkin *oma yksityinen U*, voi kokonaisuus silti maksimoida *yhteistä hyvää*: valitaan yhteiset (liikenne)säännöt siten, että yhteinen hyvä on kannattavin vaihtoehto jokaiselle. Tätä on esim. talous- ja yhteiskuntatieteissä käytetty ns. *mekanismisuunnittelu (mechanism design)*.

A_{ert} ei toisaalta saa jäädä odottelemaan *epäolennaista* informaatiota: jos taustapeilistä näkyy toinen auto, mutta se ei ole vielä laittanut ohitusvilkkua päälle, ei jäädä odottelemaan sitä, vaan ohitetaan itse.

Tekoälyn tehtäväksi nähdään nyt *tietokoneen ohjelmoiminen laskemaan funktio* $A_{\mathsf{ert}}(H_{\mathsf{nyt}}) = T_{\mathsf{nyt}}$.

```
function SKELETON-AGENT(percept) returns action
   static: memory, the agent's memory of the world

   memory ← UPDATE-MEMORY(memory, percept)
   action ← CHOOSE-BEST-ACTION(memory)
   memory ← UPDATE-MEMORY(memory, action)
   return action
```

Algoritmi 1: Älykkään toimijan perusohjelma [RN95, kuva 2.4].

Mikään ei kuitenkaan takaa, että tämä funktio (aliohjelma CHOOSE-BEST-ACTION) olisi mekaanisesti laskettavissa. Muutetaan siis tavoite *täydellisestä (perfect) laskennalliseksi (calculative)* rationaalisuudeksi, ja edelleen käytännössä *rajoitetuksi optimaalisuudeksi (bounded optimality, BO)*, jossa tätä funktiota approksimoidaan mahdollisimman hyvin ylittämättä ennalta annettuja laskentaresursseja. Toimija on siis BO-mielessä älykäs, jos se käyttää sille suodun harkinta-ajan mahdollisimman tehokkaasti, keskittyy olennaiseen.

1.2. Tekoäly ja matematiikka

Logiikka alkoi jo antiikista (luku 1.1.3), mutta vasta 1800-luvun loppupuolelta alkaen se laajeni filosofiasta (meta)matematiikkaan. Haluttiin tarkka muotoilu *todistuksen käsitteelle*: jono väitteitä. . .

1. jotka on esitetty niin täsmällisessä kielessä, ettei yhtäkään niistä voi tulkita kuin yhdellä tavalla; ja
2. jossa jonossa seuraava väitteen voidaan aina nähdä olevan välttämätön edellisten nojalla – kuten (3) väitteistä (1) ja (2).

Tämä oli keskeinen osa *David Hilbertin (1862–1943)* ohjelmaa saattaa matematiikka järkkymättömälle perustalle. Vuonna 1900 hän listasi 23 matematiikan keskeistä avointa ongelmaa, joista viimeinen oli *Entscheidungsproblem*: voidaanko *luonnollisia lukuja koskevien loogisten väitteiden totuus* ratkaista mekaanisesti?

George Boole (1815–64) oli jo 1847 yrittänyt "matematisoida logiikan" antamalla totuusarvoille laskulait lukuina 0 ja 1. Tästä syntyi myöhemmin *lauselogiikka*.

Gottlob Frege (1848–1925) kehitti 1879 *ensimmäisen kertaluvun predikaattilogiikan (First Order Predicate Calculus)*. Sen pohjalta hän yritti 1893–1903 "logisoida matematiikan". *Bertrand Russell (1872–1970)* osoitti kuitenkin 1902 menetelmässä virheen *paradoksillaan*: jos määritellään $P = \{Q : Q \notin Q\}$, niin sekä $P \in P$ että $P \notin P$ ovat mahdottomia.

Russell ja *Alfred N. Whitehead (1861–1947)* korjasivat ongelman *tyyppiteorialla*: P ei olekaan joukko, vaan "luokka". Heidän teossarjansa *Principia Mathematica (1910–13)* lienee kuuluisin yritys saavuttaa tavoite 1.

Tämän osuuden täydensi *Alfred Tarski (1902–83)* vuonna 1933 antamalla yleiset periaatteet, joilla tämän kaltaisten kielten totuus voidaan määritellä.

Tavoite 2 johti puolestaan **laskettavuuden** käsitteeseen: jo *Gottfried W. von Leibniz (1646–1716)* oli haaveillut kielestä, jossa kiistat voitaisiin ratkaista laskemalla. Piti siis määritellä intuitiivinen "laskentasäännön" tai *algoritmin* idea (joka tunnettiin jo antiikissa, ja on nimetty arabimatemaatikko *al-Khwārizmī*n mukaan, jonka *Kitāb al-jabr wa al-muqābalah* vuodelta 825 toi länsimaihin algebran). Tavoite siis oli, että jokainen todistuksen askel voitaisiin *tarkistaa* mekaanisesti.

Ensiksi vuonna 1930 *Kurt Gödel (1906–78)* edisti Hilbertin ohjelmaa huomattavasti *predikaattilogiikan täydellisyyslauseellaan*, jonka mukaan tälle kielelle oli todellakin olemassa

- todistusmenetelmä jolla voi todistaa jokaisen totuuden
- eli mekaaninen todistuksen tarkistaja K.

Tästä logiikasta puuttuu kuitenkin *induktioperiaate*, joten entscheidungsproblem jäi edelleen auki.

Gödel ratkaisi ensin tavoitteen 2 abstraktissa matemaattisessa mielessä *rekursiivisten funktioiden teoriana*. Vuonna 1936 luvussa 1.1.1 mainittu Alan Turing esitti Gödelin työn kanssa yhtäpitävän laskettavuuskäsityksen, joka kuitenkin perustui (periaatteessa) rakennettavissa oleviin mekaanisiin laitteisiin. Tämä *Turingin kone* tuli sittemmin johtavaksi tietokoneen teoreettiseksi muotoiluksi.

Sitten vuonna 1931 Gödel osoitti koko ohjelman mahdottomaksi *epätäydellisyyslauseellaan*, jonka mukaan jos kielessä L voidaan esittää luonnollisten lukujen aritmetiikka, niin siinä voidaan myös lausua tosia väitteitä, joita ei kuitenkaan voida todistaa. Näin Gödel vastasi Hilbertin 23. ongelmaan "ei".

Gödelin päättelyn idea oli, että kielellä L voidaan silloin esittää myös lainauksia sekä algoritmeja (käyttäen eräänlaista salakirjoitusta, nk.

"Gödel-numerointia"). Olkoon sitten C mikä tahansa algoritmi kielen L todistusten tarkistamiseen. Silloin kielellä L voidaan sanoa *valehtelijan paradoksi* "minä olen väite, jonka todistusta C ei pysty tarkistamaan". Turing käytti myöhemmin samaa periaatetta osoittaessaan koneidensa *pysähtymisongelman* ratkeamattomaksi.

Yritys perustaa matematiikka formaaliin logiikkaan siis kariutui. Tekoäly kuitenkin otti sen välineet käyttöönsä:

- Tietämys maailman tilasta pyritään esittämään *kuvailevasti* sitoutumatta ennalta mihinkään tiettyyn käyttöön.

  Matemaattista aksioomaa saa käyttää todistuksessa sen oletuksista johtopäätökseen tai sen johtopäätöksen vastakohdasta sen oletusten vastakohtaan.
- Tietämyksen käyttö on erillisen *päättelyjärjestelmän* reviiriä.

  Ratkaisu aksiooman käyttösuunnasta tehdään vasta todistushetkellä, ei vielä sitä kirjoitettaessa.
- Päättelyjärjestelmä käsittelee *merkkejä, ei merkityksiä*.

  Muutenhan kuulija voisi torjua todistuksen sopimattomana hänen omaan merkitysteoriaansa.

"Älykäs päättely" on silloin käsitteellinen ristiriita. . .

1960-luvun puolivälistä alkaen on teoriassa tietojenkäsittelyssä keskitytty laskettavuuden teorian sijasta **laskennan vaativuusteoriaan**, joka pyrkii arvioimaan laskulaitteiden ajan- ja muistintarpeen kasvua syötteen koon kasvaessa. Perusajatus on, että *polynominen* kasvu voidaan vielä sietää, *eksponentiaalista* ei: koneen kasvattamisen pitäisi aina näkyä myös sen menestyksellä suorittamien laskutehtävien koon selvänä kasvuna.

Vaativuusteorian tärkein uusi käsite on *NP-täydellisyys*. Siinä missä Gödel kysyi 1930 voidaanko *predikaattilogiikkaa ollenkaan* automatisoida, *Stephen Cook* ja *Richard Karp* kysyivät 1971–72 *kuinka tehokkaasti lauselogiikka* voitaisiin automatisoida. Lauselogiikassa nimittäin jokaiselle todelle väitteelle *on olemassa* jokin lyhyt perustely, mutta sen *löytäminen nopeasti* vaikuttaa varsin vaikealta. Laskentaongelmaa sanotaan NP-kovaksi, jos sen nopeasta ratkaisusta saataisiin jalostettua nopea ratkaisu tälle niin sanotulle SATISFIABILITY- eli SAT-ongelmalle, ja NP-täydelliseksi jos sama pätee myös toisinkin päin. (Sen *todistamista*, että nämä ongelmat välttämättä vaativat eksponentiaalisen ajan, pidetään teoreettisen tietojenkäsittelytieteen ja joskus jopa koko soveltavan matematiikan tärkeimpänä avoimena ongelmana.)

Neljäs matematiikan lahja tekoälylle on **todennäköisyyslaskenta**. Luvun 1.1.4 määritelmähän maksimoi *oletusarvoa sille hyödylle*, jonka toimija A_{ert} voi teollaan T_{nyt} saavuttaa, kun U sen "arvomaailmaksi". Tämä on puolestaan *päätösteoriaa (decision theory)* [RN95, luvut 16.1 ja 16.3], jonka *John von Neumann (1903–57)* ja *Oskar Morgenstern (1902–77)* perustivat 1944 ratkaisemaan, miten tehdä päätöksiä sattumanvaraisessa maailmassa.

Tarkastellaan esimerkiksi arpajaisia, jossa on jaossa yhteensä n arpaa á c $, joista yksi voittaa y $ ja muut jäävät ilman. Ostaisiko luvun 1.1.4 A_{ert} arvan, kun hän alussa omistaa x $?

Toimijan A_{ert} *omaisuuden* odotusarvo olisi päätöksestä riippuen joko

$$\begin{aligned}
E_{\text{ostaa}} &= \frac{1}{n}(x - c + y) + \frac{n-1}{n}(x - c) \\
&= x + \left(\frac{y}{n} - c\right) \ \$ \text{ tai} \\
E_{\text{ei osta}} &= x \ \$.
\end{aligned}$$

Arpajaisten järjestäjä on varmaankin valinnut $y < cn$, muutenhan niistä jäätäisiin tappiolle – joskin arvat kyllä kävisivät kaupaksi. Siis

$$E_{\text{ostaa}} < E_{\text{ei osta}}$$

joten A_{ert} ei milloinkaan sortuisi ostamaan arpaa.

Kuitenkin arvat käyvät kaupaksi; ovatko niiden ostajat siis (aina) *epä*rationaalisia toimijoita?

Eivät välttämättä: A_{ert} saattaa (perustellusti) olla sitä mieltä, että vaikka c \$ tappio *hiukan* harmittaisikin, niin voittaa y \$ tuntuisi silti *paljon* paremmalta kuin turvallinen nykytilanne x \$. Eli

$$U(x-c) \underset{\text{hiukan}}{<} U(x) \underset{\text{paljon}}{<} U(x-c+y).$$

Toimijan A_{ert} *onnellisuuden* odotusarvo onkin nyt

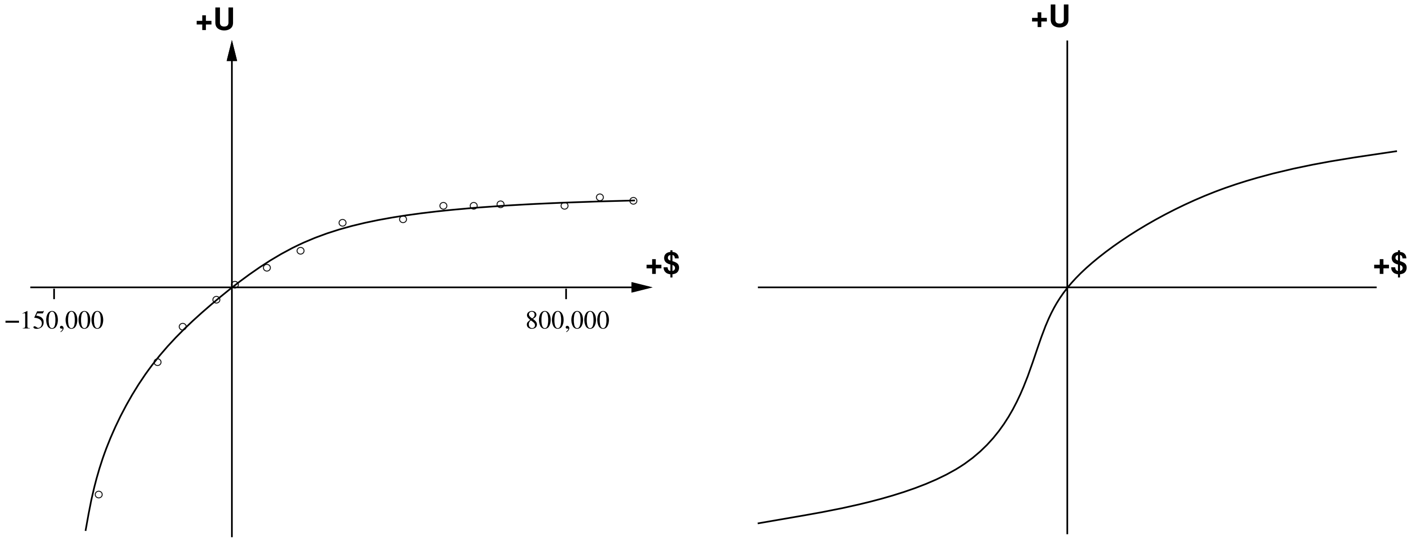
$$\begin{aligned} EU_{\text{ostaa}} &= \frac{1}{n}U(x-c+y) + \frac{n-1}{n}U(x-c) \text{ ja} \\ EU_{\text{ei osta}} &= U(x), \text{ jolloin saattaakin olla} \\ EU_{\text{ostaa}} &> EU_{\text{ei osta}}, \end{aligned}$$

eli A_{ert} saattaisikin ostaa arvan. Tai jos A_{ert} pitää uhkapelistä, niin U sisältää termin ei vain rahalle vaan myös itse osallistumiselle.

Kuvassa 1 (a) on haastattelulla selvitetty erään herra Beardin rahaa koskeva hyötymitta

$$U(x+z) = -263,31 + 22,09\log(z + 150.000\ \$),$$

missä $-150.000\,\$ < z < +800.000\,\$$ on mahdollinen tappio tai voitto. (Onko tässä luvun 1.1.2 tilanne: "älykäs arpajaiskäyttäytyminen" tarkoittaa "toimia kuten keskiverto ihiminen"?)



(a) (b)

Kuva 1: Herra Beardin hyötykäsite [RN95, kuva 16.1.].

Herra Beard *välttää riskejä*, koska käyrä on *ylöspäin kupera* [Myr74a, luku 5.9]. Useimmat ihmiset lienevät samanlaisia.

Tällaisen käyrän omistava A_{ert} ei arpaamme ostaisi:

$$\frac{1}{n}U(x-c+y)+\frac{n-1}{n}U(x-c)<U(x) \tag{4}$$

järjestyy uudelleen muotoon

$$U(x-c+y)-U(x)<(n-1)(U(x)-U(x-c)),$$

ja *differentiaalilaskennan väliarvolauseen* [Myr74a, lause 5.4.1] mukaan löytyy $x-c<x_-<x<x_+<x-c+y$, joilla

$$\begin{aligned}(y-c)U'(x_+) &< (n-1)cU'(x_-), \text{ eli}\\ c(U'(x_-)-U'(x_+)) &< cnU'(x_-)-yU'(x_+).\end{aligned} \tag{5}$$

Koska $y<cn$, seuraa epäyhtälö (5) epäyhtälöstä

$$c(U'(x_-)-U'(x_+))<y(U'(x_-)-U'(x_+)),$$

joka puolestaan seuraa epäyhtälöstä

$$c < y \tag{6}$$

funktion U kuperuuden vuoksi.

Epäyhtälö (6) on puolestaan järkevä: miksi A_{ert} ehdoin tahdoin tuhlaisi rahaa ostamalla arvan, joka on kalliimpi kuin sillä mahdollisesti saatava voitto? Tai matemaattisesti, jos U on kasvava, niin epäyhtälö (4) pätee myös silloin kun epäyhtälö (6) ei päde.

Vastaavasti voidaan laskea, että tällainen A_{ert} uskoo sananlaskun "parempi pyy pivossa kuin kymmenen oksalla": jos A_{ert} saa valita, ottaako varmat v \$ tai osallistuuko arvontaan, jossa oletusarvoisesti voittaa *enemmän kuin* v \$, niin A_{ert} valitsee varmat v \$.

Samoin tällainen A_{ert} ottaisi *vakuutuksen*, vaikka se maksaakin enemmän kuin pelätyn vahingon oletusarvo. Tähän perustuu vakuutustoiminta: ihminen liittää kotinsa palamiseen enemmän mielipahaa kuin sen rahallinen arvo ja palamistodennäköisyys edellyttäisivät, ja on valmis maksamaan turvallisuudentunteestaan ylimääräistä. Tämä erotus on vakuutusyhtiön voittomarginaali.

Kuvassa 1 (b) on käyrää laajennettu negatiiviselle puolelle olettamalla, että jossakin vaiheessa tulee *käännepiste*, jonka jälkeen lisätappiot eivät enää paljoa lisää toimijan A_{ert} surua. Eli jos A_{ert} havaitsee, että häntä joka tapauksessa uhkaa $1,000,000$ \$ tappio ellei hän tee mitään, hän saattaa päätyä uhkarohkeaan yritykseen, joka onnistuessaan estää tappion mutta epäonnistuessaan tuplaa sen. Eikä tämä ole tyhmää tai lyhytnäköistä käyttäytymistä: jos olet uppoavan Titanicin kannella, niin *todennäköisesti* tulet kuolemaan, mutta silti sinun kannattaa yrittää *mitä tahansa* joka voisi pelastaa henkesi.

Nämä laskelmat kuitenkin edellyttävät, että A_{ert} pystyy päättelemään, mitkä teot T_{nyt} ovat tällä hetkellä mahdollisia ja miten maailma muuttuu niiden seurauksena. Toisin kuin loogisissa ideaalimaailmoissa, todellisessa maailmassa tämä tietämys ei enää olekaan täsmällistä vaan ainoastaan kerätyista havainnoista H_{nyt} nähtäviä säännönmukaisuuksia. Näin päädytään *Bayesilaiseen tilastolliseen päättelyyn*, jonka periaatteet julkaistiin sen kehittäjän *Thomas Bayesin (1702-61)* kuoltua 1763. Tämä teoria selittää, miten A_{ert} voi päivittää omien maailmaa koskevien *uskomustensa voimakkuusasteita (degree of belief)* uuden tiedon karttuessa.

Esimerkki tällaisesta päättelytavasta on luku 1.1.2: hypoteesit (4a) ja (4b) vahvensivat toinen toistaan, vaikkei kumpaakaan todistettu aukottomasti loogisessa tai matemaattisessa mielessä.

Keskeinen tutkimusongelma onkin:

Miten voitaisiin parhaiten yhdistää logiikka ja Bayesin teoria?

Ratkaisua odotellessa tiivistetään toimijalähtöinen lähestymistapamme tekoälyyn seuraavasti:

- Toimijan A_{ert} perimmäiset tavoitteet on annettu ulkoa suorituskykymittana U, jonka oletusarvoa sen tulisi maksimoida teoillaan T_{nyt} päätösteorian mielessä.

  A_{ert} saattaa "tietää" oman mittansa U ja/tai rakentaa itselleen välitavoitteita johon pyrkiä.

  Mitta U vastaa, kun A_{ert} kysyy itseltään "Mitä minä oikeastaan *tahtoisin* saavuttaa?"
- A_{ert} muodostaa "maailmankuvansa" saamiensa havaintojen H_{nyt} pohjalta. Se kertoo, mitä eri vaihtoehtoiset teot T_{nyt} ovat aiemmin saaneet aikaan, ja millä todennäköisyyksillä. Tämä informaatio tarvitaan mitan U oletusarvojen laskemiseen ja vertailuun.

  Tämä kuva koostuu tilastollisesti havaituista säännönmukaisuuksista, joiden "uskottavuutta" päivitetään Bayesilaisittain havaintoaineiston karttuessa.

  Maailmankuvalla saattaa olla eksplisiittinen esitys, mutta ei välttämättä.

  Maailmankuva vastaa kysymykseen "Mitä minä *voisin* nyt saavuttaa?"

- A_{ert} joutuu päättelemään nykyisen maailmankuvansa perusteella (esimerkiksi) mitkä vaihtoehdot T_{nyt} ovat tällä hetkellä mahdollisia, mitkä aiemmin asetetut välitavoitteet ovat vielä kesken. . . Tähän yhden tavan antaa formaali logiikka, jossa kuitenkin (yleensä) unohdetaan todennäköisyydet ja uskomusten asteet.

  Tämä komponentti tarjoaa *muistin*, jonka sisällön käsittelytapoja ei välttämättä tarvitse kiinnittää etukäteen suunnittelupöydällä.

  Tämä kurssi keskittyy tämän komponentin menetelmiin:

  – symboliseen ja loogiseen tietämyksen esittämiseen
  – etsintään ja päättelyyn tästä tietämyksestä.

  Muille kursseille jäävät esimerkiksi

  – oppiminen havainnoista (koneoppinen)
  – tilastollinen päättely (kolme käsitettä -kurssit)
  – toiminta päättelyn pohjalta (robotiikka).

1.3. Tekoäly ja filosofia [RN95, luku 1.2]

Sana "teko*äly*" viittaa filosofiassa pohdittuihin ongelmiin. Tehdään siis lyhyt katsaus niihin ajatusrakennelmiin, jotka ovat edesauttaneet alan syntyä.

Ajattelua ja päättelyä on tutkittu jo antiikissa, esimerkiksi Aristoteleen syllogismeja (luku 1.1.3). Tällöin syntyi (?) ajatus siitä, että sääntöjoukko voi ohjata ihmisen mieltä – ainakin osittain; Aristoteleen ajattelussa oli mukana myös intuitiivinen äly ja tekijän sanaton tieto.

Mikä siis on oman mielemme ja meitä ympäröivän fysikaalisen maailman välinen suhde? Päädytään *ontologiaan*, filosofian alueeseen joka kysyy "mitä *pohjimmiltaan* on olemassa?"

- Jos sekä mieli että maailma ovat kumpikin olemassa erillään, päädytään *dualismiin*, jonka tunnetuin edustaja lienee ollut *René Descartes (1596–1650)*, latinaksi *Cartesius*.

  Mutta miten mieli voi silloin vaikuttaa fyysiseen maailmaan – ohjata sitä syy-seuraus -suhteilla?
- Jos mielikin nähdään fyysisen maailman tuloksena, päädytään *materialismiin*, jonka varhainen muotoilija oli Leibniz (luku 1.2).

  Mutta miten silloin voidaan selittää se, että ihmisellä näyttää olevan vapaa tahto, jolla hän voi valita mitä tehdä, eikä vain mennä luonnonlakien mukaan kuin putoava kivi?

  Usein kuulee väitettävän, että nykyinen (luonnon)tieteellinen maailmankuva olisi materialistinen. Näin pitkälle ei voine yleistää; lähinnä uskotaan, että *ainakin* fysikaalinen maailma on varmasti olemassa. Se, voidaanko siitä saada (varmaa) *tietoa*, onkin jo vaikeampi kysymys. . .

  Tekoälytutkimuksen yleisin lähtökohta lieneekin suunnilleen, että *mieli on informaation käsittelyä, aivot taas eräs fysikaalinen laite jolla tätä käsittelyä voidaan suorittaa.*

- Näiden kahden ääripään väliin sijoittuu monen tekoälyä arvostelevan filosofin omaksuva kanta, jonka mukaan vaikka mielellä onkin fysikaalinen perusta ("Brains Cause Minds", sanoo J. Searle), *ei se silti ole selitettävissä* pelkästään palauttamalla sen operaatiot fysikkaan.

  Voi esimerkiksi hyväksyä, että yksittäiset hermosolut toimivat fysiikan lakien mukaan, mutta väittää silti, ettei niiden *yhteistoiminnan tulosta* voi enää ymmärtää fysiikan lakien kautta. (Fysiikassakin ajatellaan yksittäisten hiukkasten noudattavan deterministisiä laskulakeja, mutta suurina määrinä siirrytäänkin puhumaan tilastollisesti entropiasta.)

- Filosofiassa on esitetty myös *idealistisia* kantoja, kuten *George Berkeleyn (1685–1753)* kanta, että on vain ääretön henki (= Jumala) ja äärellisiä henkiä (= ainakin ihmismielet, entä muut?), ja "fysikaalinen maailma" onkin pelkkiä ideoita, joita näillä hengillä on [Saa85, sivut 195–208].

  Helposti ajatellaan, että nämä kannat kuuluvat "selvästi arkikokemuksen vastaisina" jo menneisyyteen. Mutta itse asiassa niiden vaikutus näkyy edelleen: miten mieli saa kaipaamaansa informaatiota ulkomaailmasta, ellei tällaisina "ideoina" – ja koskeeko saatu informaatio sitten ulkomaailmaa itseään vai mielemme ideoita siitä?

Näin päädytään *tieto-oppiin (epistemologiaan)*, joka kysyy "Miten saamme informaatiota/tietoa? Onko (periaatteessakaan) mahdollista saada varmaa tietoa? Mikä on *synnynnäisten ideoiden* osuus ymmärryksessämme, vai onko mieli syntyessään kuin tyhjä taulu (Tabula Rasa) johon kokemus kirjoittaa paitsi yksittäiset havainnot myös niiden säännönmukaisuudet?"

Miten esimerkiksi saisimme tietää että "kaikki korpit ovat mustia"? Kärjistäen

Empiristi: "Kun näemme korpin toisensa jälkeen, ja ne kaikki ovat mustia, tämä sääntö alkaa tuntua yhä vahvemmin ja vahvemmin todelta."

(Tämä argumentti on sukua Bayesin teorialle luvusta 1.2. Varhaisia empiristejä olivat Berkeleyn lisäksi *Francis Bacon (1561–1626)*, *John*

Locke (1632–1704) – Tabula Rasan ehdottaja – ja *David Hume (1711–76)*.)

Rationalisti: "Mutta näin et saa *varmaa* tietoa! Voihan toisella puolen maapalloa olla valkea korppi, jota et koskaan tule nähneeksi. Tai voi olla *periaattessa* mahdollista, että sellainen syntyisi.

Ja sitäpaitsi sinulla pitää olla jo jokin *korppiuden käsite* valmiina, jotta tietäisit mitä katsoa.

Parempi on todistaa tämä sääntö loogisesti tuosta käsitteestä."

Empiristi: "Ja mitenhän sinä tuon käsitteesi mieleesi sait? Et ainakaan oppimalla, siinähän joutuisi tarkkailemaan lintuja! Vai oletko muka syntynyt kaikki mahdolliset käsitteet valmiina päässäsi?"

⋮

Immanuel Kant (1724–1804) alkoi sovitella näitä äärikantoja yhteen esittämällä, että ihmisellä on synnynnäisiä *kategorioita* – kuten syy-seuraus -suhteen käsite – joiden avulla havaintoja maailmasta voidaan alkaa jäsentämään [Saa85, sivut 227–54]. Nykykielellä [Pin97] sama ajatus kuuluisi "evoluutio on ohjannut ihmisaivoille sellaisen rakenteen, että ne käsittelevät havaintojen informaatiota sellaisilla tietyillä tavoilla, jotka synnyttävät mielen".

Tekoälyn toimijamääritelmässämme (luku 1.1.4) tämä ajatus näkyy siinä, että me suunnittelemme toimijamme perusrakenteen – eli ohjelmoimme sen, eli asetamme sen kategoriat – ennen kuin päästämme sen itseohjautuvasti maailmaan oppimaan havainnoistaan. Tekoäly ei siis ole *puhtaasti* empirismiä. . .

. . . mutta lähes: tämän vuosisadan alkupuolen metamatematiikan (luku 1.2) vastine filosofiassa oli *looginen positivismi*, joka myöhemmin lieveni *loogiseksi empirismiksi* ja loi pohjaa tämän vuosisadan anglosaksiselle analyyttiselle filosofialle. (Esimerkiksi Tarski ja Gödel tekivät sekä metamatematiikkaa että filosofiaa.) Tällä koulukunnalla oli 2 pääteesiä:

Todennettavuusteesi: Annetun väitelauseen *merkitys* (myöhemmin:

merkityksellisyyden/ymmärtämisen ehto) = sen *todentamisen menetelmä*.

Käännettävyysteesi: Jokainen merkityksellinen väitelause voidaan *kääntää* sellaiselle "puhtaalle" kielelle, jossa puhutaan ainoastaan havainnoista.

Turingin testi -määritelmä tekoälylle (luku 1.1.1) on tällainen: väitteen "kone M on älykäs" merkitykseksi annetaan koejärjestely, jossa tuomari C havainnoi syytettyjä M ja ihminen. Väite "tämä kone on älykäs" kääntyy puolestaan muotoon "tämä kone selviäisi tuosta koejärjestelystä".

Myös *Einsteinin suhteellisuusteoria* oli loogisen positivismin motivoimaa: miltä maailma näyttää, jos *havaintojen tekonopeudelle* oletetaan yläraja, ja sitä lähestytään?

Ilman käännettävyysteesiä älykkäillä toimijoillamme A_{ert} olisi *jo periaatteessa* rajoittunut ymmärrys: silloin A_{ert} voisi kohdata (tai joutua muodostamaan) käsitteen, jonka merkitystä se ei tuntisi "kategorioidensa" eikä havaintojensa pohjalta. Eikä sillä ole muuta!

Valitessamme tekoälyn määritelmäksemme älykkäät toimijat luvussa 1.1.4 teimme samalla filosofisen valinnan siitä, mikä *ohjaa* toimintaa – tässä luvussahan on vasta puhuttu siitä, mitä on olemassa, ja miten siitä saadaan tietoa. Valitsimme *utilitarismin*, jonka filosofiassa muotoili *John Stuart Mill (1806–73)* vuonna 1863.

Sen mukaan on aina syytä pyrkiä "maksimoimaan hyötyä". Tämä "hyöty" *ei* kuitenkaan välttämättä tarkoita seuraavaa:

- Sitä, että aina olisi yksikäsitteinen hyödyn mitta kuten raha luvussa 1.2.
- Sitä, että tämä periaate *selittäisi* toiminnan: kuva 1 antaa kyllä herra Beardin hyötyfunktion, mutta ei kerro *mistä syystä* se on muodostunut sellaiseksi kuin on. Sen pohjalta voi vain ennustaa hänen tulevaa käytöstään.

- Tai sitä, että A_{ert} aina maksimoisi vain *omaa* etuaan. Se voi maksimoida myös yhteistä etua – tai jopa maksimoimalla omaa etuaan maksimoida samalla yhteistä etua.

  Esimerkkinä olkoot sosiaaliset hyönteiset kuten mehiläiset. Pesässä valtaosa naaraista ($=$ työläiset) jättävät vapaaehtoisesti lisääntymisen kuningattarelle, ja keskittyvät hoitamaan sen jälkeläisiä.

  Tämä näyttää olevan vastaesimerkki "itsekkäälle evoluutioteorialle", jonka mukaan jokainen eläin pyrkii lisääntymään, koska sen jokainen geeni pyrkii kopioitumaan.

  Mehiläiset ovat kuitenkin *haplodiploideja*:

	koiras	**naaras**
ihminen	äidin X, isän Y	äidin X, isän X
mehiläinen	kuningattaren X, kuhnurin Y	kuningattaren X

  Taulukko 1: Ihmisen ja mehiläisen sukukromosomit

  Työläinen hoitaa siis tulevaa kaksossisartaan hoitaessaan kuningattaren naarasmunaa. Se taas on yhtä tehokasta kopiointia kuin oma lisääntyminen. Kuningattaren koirasmunan hoitamiseen kuluvasta vaivasta vain puolet menee työläisen oman X-kromosomin kopiolle, ja toinen puoli isäkuhnurin Y-kopiolle. Silloin työläisen kannattaa jakaa aikansa suhteessa $3 : 1$ naaras- ja koirasmunien välillä.

  Kuningattarelle puolestaan naaras- ja koirasmunat ovat yhtä arvokkaita, kun lasketaan niiden kopiointitehokkuutta lapsenlapsituotannossa, joten hänen näkökulmastaan ajankäyttö saisi olla $1 : 1$.

  Todellisessa pesässä työläisten ajankäyttö vaihteleekin tuolla välillä riippuen siitä, ovatko työläiset vai kuningatar niskan päällä.

  Näin siis usean itsekkään toimijan yhteisössä voi syntyä (ulkoapäin katsoen) epäitsekästä yhteistyötä, kuten luvussa 1.1.4 jo todettiin.

Viitteet

[AK83] J.F. Allen and J.A. Koomen. Planning using a temporal world model. In *International Joint Conference on Artificial Intelligence*, pages 741–747, 1983.

[AS96] H. Abelson and G.J. Sussman. *Structure and Interpretation of Computer Algorithms*. The MIT Press, second edition, 1996.

[Bod90] M.A. Boden, editor. *The Philosophy of Artificial Intelligence*. Oxford Readings in Philosophy. Oxford University Press, 1990.

[Car97] B. Carpenter. *Type-Logical Semantics*. The MIT Press, 1997.

[Cho97] N. Chomsky. B. F. Skinnerin teos *verbal behaviour*. In P. Raatikainen, editor, *Ajattelu, kieli, merkitys*, pages 279–310. Gaudeamus, 1997. Suom. R. Vilkko.

[CL73] C-L. Chang and R.C-T. Lee. *Symbolic Logic and Mechanical Theorem Proving*. Academic Press, 1973.

[CR91] W. Clinger and J. Rees. Revised^4 report on the algorithmic language Scheme, 1991.

[CSC90] P.M. Churchland and P. Smith Churchland. Could a machine think? *Scientific American*, pages 26–31, 1990.

[Dam99] A. Damasio. *The Feeling of What Happens*. William Heinemann, 1999.

[DD88] H.L. Dreyfus and S.E. Dreyfus. Making a mind versus modelling the brain: Artificial intelligence back at a branch-point. *Artificial Intelligence*, 1988.

[Dea98] T.W. Deacon. *The Symbolic Species: The Co-Evolution of Language and the Human Brain*. Penguin Books, 1998.

[Myr74a] L. Myrberg. *Differentiaali- ja integraalilaskenta*, volume 1. Kirjayhtymä, 1974.

[Myr74b] L. Myrberg. *Differentiaali- ja integraalilaskenta*, volume 2. Kirjayhtymä, 1974.

[Nil98] N.J. Nilsson. *Artificial Intelligence: A New Synthesis*. Morgan-Kaufmann, 1998.

[Pen89] R. Penrose. *The Emperor's New Mind*. Oxford University Press, 1989.

[Pin94] S. Pinker. *The Language Instinct*. William Morrow and Company, Inc., 1994.

[Pin97] S. Pinker. *How the Mind Works*. W. W. Norton & Company, Inc., 1997.

[RN95] S. Russell and P. Norvig. *Artificial Intelligence: a Modern Approach*. Prentice-Hall, 1995.

[Saa85] E. Saarinen. *Länsimaisen filosofian historia huipulta huipulle*. Werner Söderström oy, 1985.

[Sea80] J. Searle. Minds, brains, and programs. *The Behavioral and Brain Sciences*, 3:417–424, 1980. [Bod90, §3].

[Sea90] J. Searle. Is the brain's mind a computer program? *Scientific American*, pages 20–25, 1990.

[TE88] T. Tononi and G.M. Edelman. Consciousness and complexity. *Science*, pages 1846–1851, 1988.

[Tur50] A.M. Turing. Computing machinery and intelligence. *Mind*, LIX(2236):433–460, 1950. [Bod90, §2].