Hvem er smartest, menneske eller maskin? Betydningen av Gödels ufullstendighetsteorem

Øystein Linnebo

Filosofi – Universitet i Oslo

HON 1000, 29. september 2021

Hvem er smartest, menneske eller maskin?

Hvem er smartest, menneske eller maskin?

Descartes (1596–1650): Bare mennesker har fornuft og kreativitet!

Kunstig intelligens viser at datamaskiner på mange vis er mer intelligente enn mennesker.

Kunstig intelligens viser at datamaskiner på mange vis er mer intelligente enn mennesker.

Er mennesker mer intelligente enn datamaskiner på andre vis?

Kunstig intelligens viser at datamaskiner på mange vis er mer intelligente enn mennesker.

Er mennesker mer intelligente enn datamaskiner på andre vis?

- Sosial intelligens, emosjonell intelligens, etc.

Kunstig intelligens viser at datamaskiner på mange vis er mer intelligente enn mennesker.

Er mennesker mer intelligente enn datamaskiner på andre vis?

- Sosial intelligens, emosjonell intelligens, etc.
- Er vi mer intelligente enn datamaskiner *selv på datamaskinenes egen banehalvdel*, nemlig matematikk?

Either . . . the human mind . . . infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable [mathematical] problems. (Gödel, Gibbs lecture, 1951)

Either . . . the human mind . . . infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable [mathematical] problems. (Gödel, Gibbs lecture, 1951)

"Optimisme" vs. "pessimisme"

Matematiske problemer

Ethvert naturlig tall større enn 1 kan på entydig måte skrives som et produkt av primtall.

Matematiske problemer

Ethvert naturlig tall større enn 1 kan på entydig måte skrives som et produkt av primtall.

$$18 = 2 \times 3^2$$

$$57 = 3 \times 19$$

$$1,050 = 2 \times 3 \times 5^2 \times 7$$

"Fundamental Theorem of Arithmetic"

Det er lett å se at $3^2 + 4^2 = 5^2$.

Det er lett å se at $3^2 + 4^2 = 5^2$.

Fermat's Last Theorem (1637?, 1994)

Det finnes ingen naturlige tall a, b, c unntatt 0 slik at:

$$a^3 + b^3 = c^3.$$

Og likedan for høyere eksponenter.

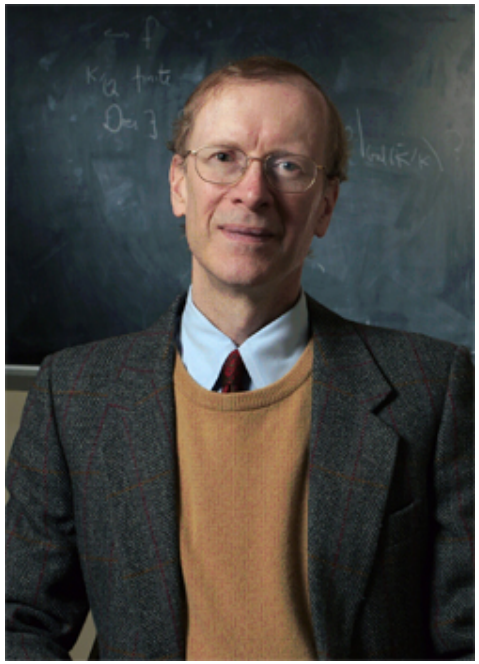
Det er lett å se at $3^2 + 4^2 = 5^2$.

Fermat's Last Theorem (1637?, 1994)

Det finnes ingen naturlige tall a, b, c unntatt 0 slik at:

$$a^3 + b^3 = c^3.$$

Og likedan for høyere eksponenter.



Goldbach's conjecture (1742)

Ethvert partall større enn 2 kan skrives som en sum av to primtall.

Goldbach's conjecture (1742)

Ethvert partall større enn 2 kan skrives som en sum av to primtall.

$$4 = 2 + 2$$

$$6 = 3 + 3$$

$$8 = 5 + 3$$

$$\ldots$$

$$84 = 79 + 5$$

$$\ldots$$

Hvordan løser vi et matematisk problem?

Hvordan løser vi et matematisk problem?

Problemet kan ha et *positivt* eller *negativt* svar.

Hvordan løser vi et matematisk problem?

Problemet kan ha et *positivt* eller *negativt* svar.

Svaret *bevises* fra matematiske aksiomer (grunnprinsipper).

David Hilbert (1862–1943)

David Hilbert (1862–1943)

> *We must not believe those, who today, with philosophical bearing and deliberative tone, prophesy the fall of culture and accept the ignorabimus. For us there is no ignorabimus, and in my opinion none whatever in natural science. In opposition to the foolish ignorabimus our slogan shall be: Wir müssen wissen — wir werden wissen. (1930)*

Hvor mye matematikk kan en datamaskin utføre?

En datamaskin er programmert til å forholde seg til en gitt mengde A av matematiske aksiomer.

Hvor mye matematikk kan en datamaskin utføre?

En datamaskin er programmert til å forholde seg til en gitt mengde A av matematiske aksiomer.

Kan disse matematiske aksiomene gi et svar på *alle* matematiske problemer?

Hvor mye matematikk kan en datamaskin utføre?

En datamaskin er programmert til å forholde seg til en gitt mengde A av matematiske aksiomer.

Kan disse matematiske aksiomene gi et svar på *alle* matematiske problemer?

Gödels første ufullstendighetsteorem (1931)

Såfremt aksiomene i A ikke er selvmotsigende, kan *ikke* alle tallteoretiske problemer løses med utgangspunkt i disse aksiomene.

Hvordan klarte Gödel å bevise dette?

Løgnerparadokset:

(1) Denne setningen er usann.

Hvis den er sann, er den usann. Men hvis den er usann, er den sann!

Hvordan klarte Gödel å bevise dette?

Løgnerparadokset:

(1) Denne setningen er usann.

Hvis den er sann, er den usann. Men hvis den er usann, er den sann!

Gödel-setningen:

(2) Denne setningen kan ikke bevises utfra grunnprinsippene i A.

Hvordan klarte Gödel å bevise dette?

Løgnerparadokset:

(1) Denne setningen er usann.

Hvis den er sann, er den usann. Men hvis den er usann, er den sann!

Gödel-setningen:

(2) Denne setningen kan ikke bevises utfra grunnprinsippene i A.

Hvis denne setningen er usann, så kan den bevises fra A. Men vi kan ikke bevise en usannhet. Følgelig må setningen være sann og dermed ikke bevisbar fra A!

Hvordan klarte Gödel å bevise dette?

Løgnerparadokset:

(1) Denne setningen er usann.

Hvis den er sann, er den usann. Men hvis den er usann, er den sann!

Gödel-setningen:

(2) Denne setningen kan ikke bevises utfra grunnprinsippene i A.

Hvis denne setningen er usann, så kan den bevises fra A. Men vi kan ikke bevise en usannhet. Følgelig må setningen være sann og dermed ikke bevisbar fra A!

Gitt en vilkårlig mengde aksiomer A, har *vi* altså bevist noe som ikke kan bevises i A.

Hvordan klarte Gödel å bevise dette?

Løgnerparadokset:

(1) Denne setningen er usann.

Hvis den er sann, er den usann. Men hvis den er usann, er den sann!

Hvordan klarte Gödel å bevise dette?

Løgnerparadokset:

(1) Denne setningen er usann.

Hvis den er sann, er den usann. Men hvis den er usann, er den sann!

Gödel-setningen:

(2) Denne setningen kan ikke bevises utfra grunnprinsippene i A.

Hvordan klarte Gödel å bevise dette?

Løgnerparadokset:

(1) Denne setningen er usann.

Hvis den er sann, er den usann. Men hvis den er usann, er den sann!

Gödel-setningen:

(2) Denne setningen kan ikke bevises utfra grunnprinsippene i A.

Hvis denne setningen er usann, så kan den bevises fra A. Men vi kan ikke bevise en usannhet. Følgelig må setningen være sann og dermed ikke bevisbar fra A!

Hvordan klarte Gödel å bevise dette?

Løgnerparadokset:

(1) Denne setningen er usann.

Hvis den er sann, er den usann. Men hvis den er usann, er den sann!

Gödel-setningen:

(2) Denne setningen kan ikke bevises utfra grunnprinsippene i A.

Hvis denne setningen er usann, så kan den bevises fra A. Men vi kan ikke bevise en usannhet. Følgelig må setningen være sann og dermed ikke bevisbar fra A!

Gitt en vilkårlig mengde aksiomer A, har *vi* altså bevist noe som ikke kan bevises i A.

Hadde Descartes delvis rett likevel?

Så for enhver mengde A av matematiske grunnprinsipper (som ikke er selvmotsigende) finnes det matematiske påstander som ikke lar seg bevise fra A—men som *vi mennesker* kan "se" at er sanne.

Hadde Descartes delvis rett likevel?

Så for enhver mengde A av matematiske grunnprinsipper (som ikke er selvmotsigende) finnes det matematiske påstander som ikke lar seg bevise fra A—men som *vi mennesker* kan "se" at er sanne.

Så vi kan løse matematiske problemer som en datamaskin ikke kan løse!

Hadde Descartes delvis rett likevel?

Så for enhver mengde A av matematiske grunnprinsipper (som ikke er selvmotsigende) finnes det matematiske påstander som ikke lar seg bevise fra A—men som *vi mennesker* kan "se" at er sanne.

Så vi kan løse matematiske problemer som en datamaskin ikke kan løse!

Så vi er ikke bare biologiske maskiner . . . ?

Hadde Descartes delvis rett likevel?

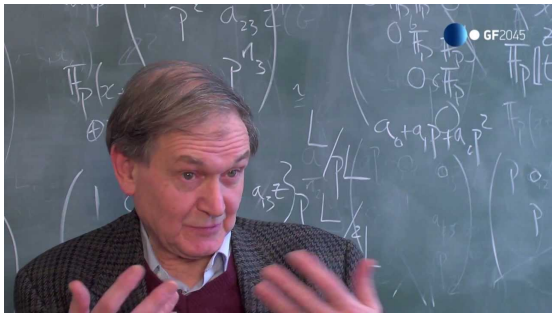
Så for enhver mengde A av matematiske grunnprinsipper (som ikke er selvmotsigende) finnes det matematiske påstander som ikke lar seg bevise fra A—men som *vi mennesker* kan "se" at er sanne.

Så vi kan løse matematiske problemer som en datamaskin ikke kan løse!

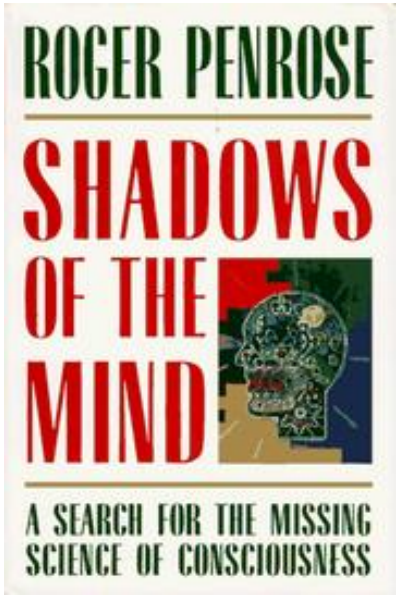
Så vi er ikke bare biologiske maskiner . . . ?

Så Descartes hadde rett likevel . . . ???

John Lucas (1929–2020)



Roger Penrose (1931–)

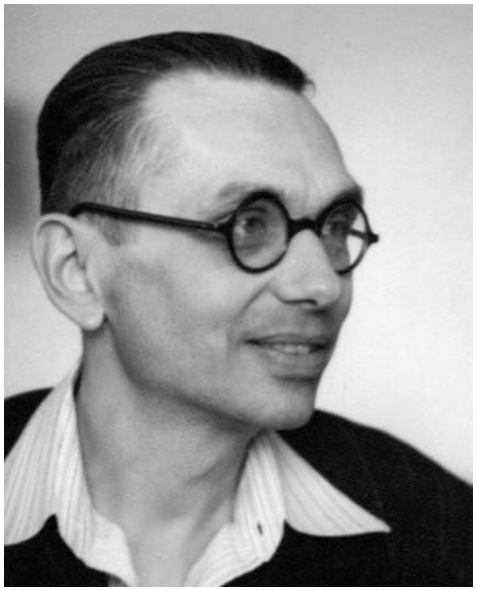


Eller kanskje en logisk eller filosofisk feil har sneket seg inn i argumentet?

Eller kanskje en logisk eller filosofisk feil har sneket seg inn i argumentet?

Gödel selv trekker en forholdsvis forsiktig konklusjon:

> *Either . . . the human mind . . . infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable [mathematical] problems. (Gödel, Gibbs lecture, 1951)*



La oss se nærmere på Gödels ufullstendighetsteorem . . .

La oss se nærmere på Gödels ufullstendighetsteorem . . .

. . . samt på forholdet mellom “det menneskelige sinn” og “en endelig maskin”

La oss se nærmere på Gödels ufullstendighetsteorem . . .

. . . samt på forholdet mellom “det menneskelige sinn” og “en endelig maskin”

. . . og begrepet om “absolutt bevisbarhet”.

A key step was to show that arithmetic can:

- speak of its own language;
- speak of its own principles;
- speak of what can and cannot be proved from its own principles.

Annie Bonnie is tall not

Annie	Bonnie	is	tall	not
1	3	5	7	9

Annie	Bonnie	is	tall	not
1	3	5	7	9

Annie is tall:

Annie	Bonnie	is	tall	not
1	3	5	7	9

Annie is tall: $2^1 \times 3^5 \times 5^7$

Annie	Bonnie	is	tall	not
1	3	5	7	9

Annie is tall: $2^1 \times 3^5 \times 5^7$

Bonnie is not tall:

Annie	Bonnie	is	tall	not
1	3	5	7	9

Annie is tall: $2^1 \times 3^5 \times 5^7$

Bonnie is not tall: $2^3 \times 3^5 \times 5^9 \times 7^7$

Diagonal-lemmaet

La T være en teori som inneholder tilstrekkelig aritmetikk. Da kan vi for enhver formel $B(y)$ finne en setning φ slik at vi i T kan bevise:

$$\varphi \leftrightarrow B(\ulcorner \varphi \urcorner)$$

Diagonal-lemmaet

La T være en teori som inneholder tilstrekkelig aritmetikk. Da kan vi for enhver formel $B(y)$ finne en setning φ slik at vi i T kan bevise:

$$\varphi \leftrightarrow B(\ulcorner \varphi \urcorner)$$

Så φ "sier": jeg er B!

Diagonal-lemmaet

La T være en teori som inneholder tilstrekkelig aritmetikk. Da kan vi for enhver formel $B(y)$ finne en setning φ slik at vi i T kan bevise:

$$\varphi \leftrightarrow B(\ulcorner \varphi \urcorner)$$

Så φ "sier": jeg er B!

Vi kan med andre ord *simulere* selv-referanse i aritmetikk.

Theorem (Gödels første ufullstendighetsteorem)

La T være en konsistent teori som inneholder tilstrekkelig aritmetikk. Da finnes en setning G_T som verken kan bevises eller motbevises i T.

Theorem (Gödels første ufullstendighetsteorem)

La T være en konsistent teori som inneholder tilstrekkelig aritmetikk. Da finnes en setning G_T som verken kan bevises eller motbevises i T.

Bevis. Vi finner en formel $Prov(x)$ som representerer at setningen x kan bevises i T. Vi anvender så diagonal-lemmaet på '$\neg Prov(x)$', hvilket gir

$$G_T \leftrightarrow \neg Prov(\ulcorner G_T \urcorner)$$

Dette gir Gödel-setningen for T, som T verken beviser eller motbeviser!

Theorem (Gödels første ufullstendighetsteorem)

La T være en konsistent teori som inneholder tilstrekkelig aritmetikk. Da finnes en setning G_T som verken kan bevises eller motbevises i T.

Bevis. Vi finner en formel $Prov(x)$ som representerer at setningen x kan bevises i T. Vi anvender så diagonal-lemmaet på '$\neg Prov(x)$', hvilket gir

$$G_T \leftrightarrow \neg Prov(\ulcorner G_T \urcorner)$$

Dette gir Gödel-setningen for T, som T verken beviser eller motbeviser!

NB! Siden G_T "sier" at den ikke kan bevises, er setningen faktisk sann!

Anta at du er en maskin som implementerer teorien T. Da kan du p.g.a. Gödels teorem ikke bevise G_T.

Anta at du er en maskin som implementerer teorien T. Da kan du p.g.a. Gödels teorem ikke bevise G_T.

Men vi "så" nettopp at G_T er sann!

Anta at du er en maskin som implementerer teorien T. Da kan du p.g.a. Gödels teorem ikke bevise G_T.

Men vi "så" nettopp at G_T er sann!

Nei! Vi beviste at dersom T er konsistent, så kan ikke T bevise G_T.

Anta at du er en maskin som implementerer teorien T. Da kan du p.g.a. Gödels teorem ikke bevise G_T.

Men vi "så" nettopp at G_T er sann!

Nei! Vi beviste at dersom T er konsistent, så kan ikke T bevise G_T.

Så dersom du er en maskin som implementerer teorien T, så kan du *ikke* bevise at T er konsistent.

Gödels disjunksjon

Hvis du er en maskin, så finnes det en teori T som du implementerer.

Gödels disjunksjon

Hvis du er en maskin, så finnes det en teori T som du implementerer.

I så fall blir Gödel-setningen til T, G_T, absolutt ubevisbar.

Gödels disjunksjon

Hvis du er en maskin, så finnes det en teori T som du implementerer.

I så fall blir Gödel-setningen til T, G_T, absolutt ubevisbar.

Alternativt er du ikke en maskin.

Gödels disjunksjon

Hvis du er en maskin, så finnes det en teori T som du implementerer.

I så fall blir Gödel-setningen til T, G_T, absolutt ubevisbar.

Alternativt er du ikke en maskin.

> *Either . . . the human mind . . . infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable [mathematical] problems. (Gödel, Gibbs lecture, 1951)*

Forholdet mellom “det menneskelige sinn” og “en endelig maskin”

Forholdet mellom "det menneskelige sinn" og "en endelig maskin"

Vi idealiserer slik at vi ser bort fra vår dødelighet, feilbarlighet, osv.

Forholdet mellom "det menneskelige sinn" og "en endelig maskin"

Vi idealiserer slik at vi ser bort fra vår dødelighet, feilbarlighet, osv.

Hvordan skal vi idealisere?

Forholdet mellom "det menneskelige sinn" og "en endelig maskin"

Vi idealiserer slik at vi ser bort fra vår dødelighet, feilbarlighet, osv.

Hvordan skal vi idealisere?

Noen ganger er det opplagt hvordan man idealiserer, f.eks. friksjonsfritt plan.

Forholdet mellom "det menneskelige sinn" og "en endelig maskin"

Vi idealiserer slik at vi ser bort fra vår dødelighet, feilbarlighet, osv.

Hvordan skal vi idealisere?

Noen ganger er det opplagt hvordan man idealiserer, f.eks. friksjonsfritt plan.

Men hvordan idealiserer vi vår egen kapasitet? Eneste mulighet pr. i dag er som en formell teori eller en datamaskin.

Forholdet mellom "det menneskelige sinn" og "en endelig maskin"

Vi idealiserer slik at vi ser bort fra vår dødelighet, feilbarlighet, osv.

Hvordan skal vi idealisere?

Noen ganger er det opplagt hvordan man idealiserer, f.eks. friksjonsfritt plan.

Men hvordan idealiserer vi vår egen kapasitet? Eneste mulighet pr. i dag er som en formell teori eller en datamaskin. —Men da bygger idealiseringen inn at den ene parten ("pessimisme") i diskusjonen vår vinner!

Hva er et matematisk aksiom?

Euklidske oppfatning av matematikk: Vi begynner med selv-innlysende aksiomer og resonnerer strengt logisk.

Hva er et matematisk aksiom?

Euklidske oppfatning av matematikk: Vi begynner med selv-innlysende aksiomer og resonnerer strengt logisk.

Denne modellen begynner å bryte sammen ca. 1900 (Russell, Gödel, Quine). Aksiomer trenger ikke være selv-innlysende, men kan også begrunnes via sine konsekvenser.

Hva er et matematisk aksiom?

Euklidske oppfatning av matematikk: Vi begynner med selv-innlysende aksiomer og resonnerer strengt logisk.

Denne modellen begynner å bryte sammen ca. 1900 (Russell, Gödel, Quine). Aksiomer trenger ikke være selv-innlysende, men kan også begrunnes via sine konsekvenser.

Selv ideelle aktører endre sin oppfatning ang. aksiomer ettersom matematikken utvikler seg. Så datamaskin-modellen for en "ideell matematiker" blir upassende.

Oppsummering

Either . . . the human mind . . . infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable [mathematical] problems. (Gödel, Gibbs lecture, 1951)

Oppsummering

> *Either . . . the human mind . . . infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable [mathematical] problems. (Gödel, Gibbs lecture, 1951)*

1. Ingenting tyder på at vi kan bevise første disjunkt ("optimisme").

Oppsummering

> *Either . . . the human mind . . . infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable [mathematical] problems. (Gödel, Gibbs lecture, 1951)*

1. Ingenting tyder på at vi kan bevise første disjunkt ("optimisme").

2. Å tenkte på det menneskelige sinnet som en maskin er uansett bare en modell.

Oppsummering

> *Either . . . the human mind . . . infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable [mathematical] problems. (Gödel, Gibbs lecture, 1951)*

1. Ingenting tyder på at vi kan bevise første disjunkt ("optimisme").

2. Å tenkte på det menneskelige sinnet som en maskin er uansett bare en modell.

- Men modellen og den matematiske teorien omkring den har vist seg fruktbare.

Oppsummering

> *Either . . . the human mind . . . infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable [mathematical] problems. (Gödel, Gibbs lecture, 1951)*

1. Ingenting tyder på at vi kan bevise første disjunkt ("optimisme").

2. Å tenkte på det menneskelige sinnet som en maskin er uansett bare en modell.

- Men modellen og den matematiske teorien omkring den har vist seg fruktbare.

3. Begrepet om "absolutt bevisbarhet" blir problematisk i lys av nyere matematikk.

Boolos, G., Burgess, J., P., R., and Jeffrey, C. (2007).

Computability and Logic.

Cambridge University Press.

Koellner, P. (2018a).

On the question of whether the mind can be mechanized, I: From Gödel to Penrose.

Journal of Philosophy, 115(7):337–360.

Koellner, P. (2018b).

On the question of whether the mind can be mechanized, II: Penrose's new argument.

Journal of Philosophy, 115(9):453–484.

Shapiro, S. (1998).

Incompleteness, mechanism, and optimism.

Bulletin of Symbolic Logic, 4(3):273–302.