Föreläsning på MVE420: Nya teknologier, global risk och mänskligheten framtid

Beslutsteori, rationalitet och mänsklig irrationalitet

13 april 2021

Olle Häggström

https://www.chalmers.se/en/Staff/Pages/olle-haggstrom.aspx

http://haggstrom.blogspot.com/

En första övning i beslutsteori

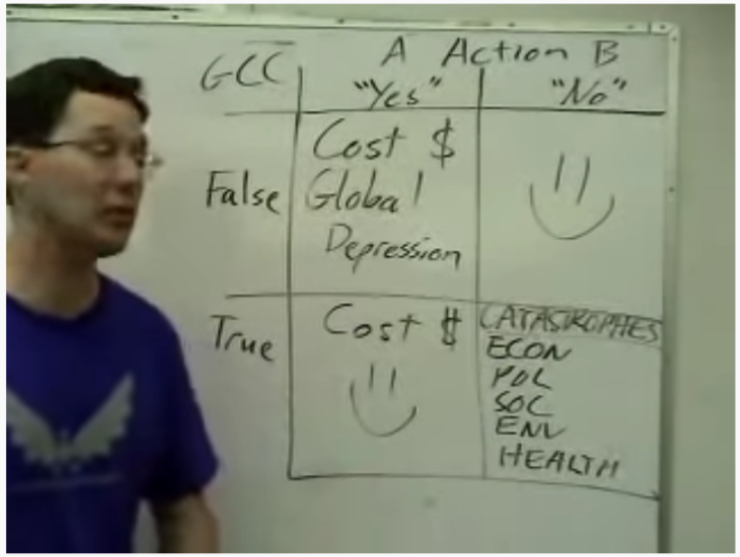
En första övning i beslutsteori

Vad bör en beslutsfattare, som har ofullständig information om världens beskaffenhet (lever vi i värld I eller värld II?), göra – välja handling A eller handling B?

	värld I	värld II
handling A	\$10	\$10
handling B	\$20	\$5

YouTube SE

Search



The Most Terrifying Video You'll Ever See

7,084,912 views • Jun 8, 2007

25K 7.1K SHARE S

En första övning i beslutsteori

Vad bör en beslutsfattare, som har ofullständig information om världens beskaffenhet (lever vi i värld I eller värld II?), göra – välja handling A eller handling B?

	värld I	värld II
handling A	\$10	\$10
handling B	\$20	\$5

En första övning i beslutsteori

Vad bör en beslutsfattare, som har ofullständig information om världens beskaffenhet (lever vi i värld I eller värld II?), göra – välja handling A eller handling B?

	värld I	värld II
handling A	\$10	\$10
handling B	\$20	\$5

Sens moral: besultsfattare behöver sannolikheter!

	värld I	värld II
handling A	\$10	\$10
handling B	\$20	\$5

Antag att värld I har sannolikhet p och värld II har sannolikhet $1 - p$.

	värld I	värld II
handling A	\$10	\$10
handling B	\$20	\$5

Antag att värld I har sannolikhet p och värld II har sannolikhet $1 - p$. Då ger...

handling A förväntat utbyte $\$(10p + 10(1 - p)) = \$10$

	värld I	värld II
handling A	\$10	\$10
handling B	\$20	\$5

Antag att värld I har sannolikhet p och värld II har sannolikhet $1 - p$. Då ger...

handling A förväntat utbyte $\$(10p + 10(1 - p)) = \$10$
handling B förväntat utbyte $\$(20p + 5(1 - p)) = \$(5 + 15p)$

	värld I	värld II
handling A	\$10	\$10
handling B	\$20	\$5

Antag att värld I har sannolikhet p och värld II har sannolikhet $1 - p$. Då ger...

handling A förväntat utbyte $\$(10p + 10(1 - p)) = \$10$
handling B förväntat utbyte $\$(20p + 5(1 - p)) = \$(5 + 15p)$

...och handling B är bättre än A om och endast om $p > \frac{1}{3}$.

Men stopp och belägg! Antag att $p = \frac{1}{2}$ (värld I och värld II lika sannolika) och att utbytet ges av...

	värld I	värld II
handling A	10 000 000 kr	10 000 000 kr
handling B	30 000 000 kr	0 kr

Men stopp och belägg! Antag att $p = \frac{1}{2}$ (värld I och värld II lika sannolika) och att utbytet ges av...

	värld I	värld II
handling A	10 000 000 kr	10 000 000 kr
handling B	30 000 000 kr	0 kr

Vad väljer ni?

Men stopp och belägg! Antag att $p = \frac{1}{2}$ (värld I och värld II lika sannolika) och att utbytet ges av...

	värld I	värld II
handling A	10 000 000 kr	10 000 000 kr
handling B	30 000 000 kr	0 kr

Vad väljer ni?

Direkt beslutsteoretisk kalkyl ger att handling A har förväntad vinst 10 000 000 kr, och handling B har förväntad vinst 15 000 000 kr. Kalkylen förordar alltså handling B.

Men stopp och belägg! Antag att $p = \frac{1}{2}$ (värld I och värld II lika sannolika) och att utbytet ges av...

	värld I	värld II
handling A	10 000 000 kr	10 000 000 kr
handling B	30 000 000 kr	0 kr

Vad väljer ni?

Direkt beslutsteoretisk kalkyl ger att handling A har förväntad vinst 10 000 000 kr, och handling B har förväntad vinst 15 000 000 kr. Kalkylen förordar alltså handling B.

Ändå väljer jag handling A!

Valet går att försvara med genom att ersätta *pengamaximering* med *nyttomaximering*. Antag att...

0 kr	svarar mot	0 utiler
10 000 000 kr	svarar mot	100 utiler
30 000 000 kr	svarar mot	150 utiler

Valet går att försvara med genom att ersätta *pengamaximering* med *nyttomaximering*. Antag att...

0 kr	svarar mot	0 utiler
10 000 000 kr	svarar mot	100 utiler
30 000 000 kr	svarar mot	150 utiler

Beslutskalkylen kan då baseras på matrisen

	värld I	värld II
handling A	100 utiler	100 utiler
handling B	150 utiler	0 utiler

Valet går att försvara med genom att ersätta *pengamaximering* med *nyttomaximering*. Antag att...

0 kr	svarar mot	0 utiler
10 000 000 kr	svarar mot	100 utiler
30 000 000 kr	svarar mot	150 utiler

Beslutskalkylen kan då baseras på matrisen

	värld I	värld II
handling A	100 utiler	100 utiler
handling B	150 utiler	0 utiler

Med $p = \frac{1}{2}$ ger handling A förväntat utbyte 100 utiler, medan handling B ger förväntat utbyte 75 utiler, varför handling A är bättre.

Att översätta pengar till nytta med hjälp av en ickelinjär nyttofunktion är ganska vanligt.

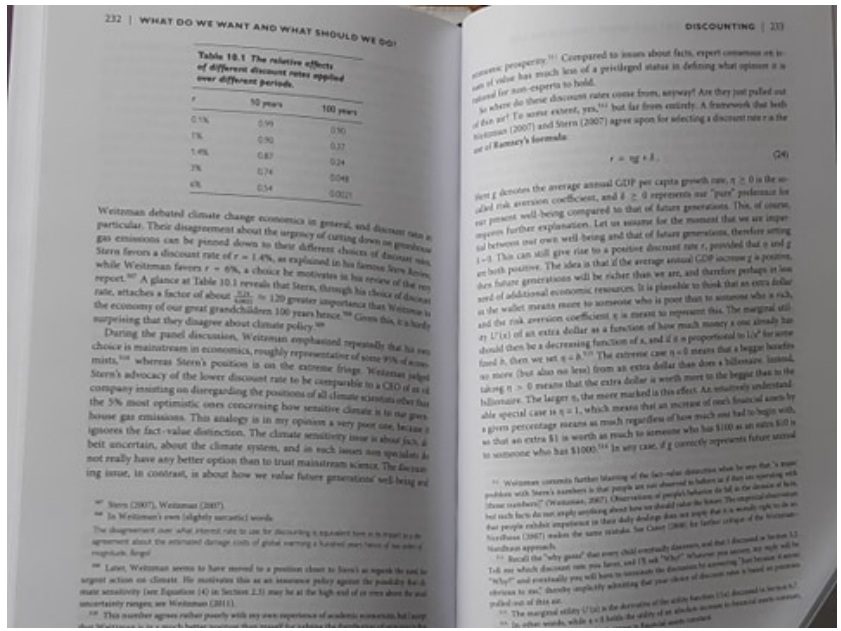
Att översätta pengar till nytta med hjälp av en ickelinjär nyttofunktion är ganska vanligt.

Eftersom en krona extra i plånboken som regel gör större skillnad för den fattige än för den rike, väljs nyttofunktionen vanligtvis konkav, som i **Ramseys formel** för tidsdiskontering.

Att översätta pengar till nytta med hjälp av en ickelinjär nyttofunktion är ganska vanligt.

Eftersom en krona extra i plånboken som regel gör större skillnad för den fattige än för den rike, väljs nyttofunktionen vanligtvis konkav, som i **Ramseys formel** för tidsdiskontering.

Frank Ramsey (1903-1930)



Here Be Dragons, s 232-233

Den första beslutsteoretiska kalkylen

Den första beslutsteoretiska kalkylen

Blaise Pascal (1623-1662)

Den första beslutsteoretiska kalkylen

Blaise Pascal (1623-1662)

Pascal's Wager

	God Exists	God Doesn't
Belief	Eternity in Heaven (+∞)	Wasted Life With False Belief (-1)
Disbelief	Eternity in Hell (-∞)	Didn't Waste Life With False Belief (+1)

Pascal's Wager

Om Guds existens har sannolikhet $\varepsilon > 0$ så ger...

tro förväntat utbyte $+\infty \cdot \varepsilon - 1(1-\varepsilon) = +\infty$

Pascal's Wager

Om Guds existens har sannolikhet $\varepsilon > 0$ så ger...

tro	förväntat utbyte	$+\infty \cdot \varepsilon - 1(1-\varepsilon) = +\infty$
otro	förväntat utbyte	$-\infty \cdot \varepsilon + 1(1-\varepsilon) = -\infty$

Pascal's Wager

Om Guds existens har sannolikhet $\varepsilon > 0$ så ger...

tro	förväntat utbyte	$+\infty \cdot \varepsilon - 1(1-\varepsilon) = +\infty$
otro	förväntat utbyte	$-\infty \cdot \varepsilon + 1(1-\varepsilon) = -\infty$

...vilket gör att **tro** är att föredra framför **otro** oavsett hur litet $\varepsilon > 0$ är.

Det finns anledning att vara försiktig med *Pascal's Wager*-liknande argument, där ytterst små sannolikheter för mycket stora utfall tillåts dominera kalkylen.

Det finns anledning att vara försiktig med *Pascal's Wager*-liknande argument, där ytterst små sannolikheter för mycket stora utfall tillåts dominera kalkylen.

Ett bekymmer för oss som studerar existentiell risk är att om man accepterar skattningar liknande dem Bostrom gjort (se min föreläsning om etik) för antalet framtida människoliv om vi inte går under –

- 10^{16} med "business as usual",
- 10^{34} med rymdkolonisering, och
- 10^{54} med uppladdning

– så hamnar man lätt i just sådana.

Ett extremt otrevligt tankeexperiment

Ett extremt otrevligt tankeexperiment

Antag att du är USA:s president, och att CIA-chefen ger dig pålitlig information om att det någonstans i Tyskland gömmer sig en terrorist som arbetar med att bygga ett domedagsvapen, som avser använda detta för att förgöra mänskligheten, och som har en chans på miljonen att lyckas med planen. Bör du attackera Tyskland med fullskaligt kärnvapenangrepp, eller bör du avstå?

Ett extremt otrevligt tankeexperiment

Antag att du är USA:s president, och att CIA-chefen ger dig pålitlig information om att det någonstans i Tyskland gömmer sig en terrorist som arbetar med att bygga ett domedagsvapen, som avser använda detta för att förgöra mänskligheten, och som har en chans på miljonen att lyckas med planen. Bör du attackera Tyskland med fullskaligt kärnvapenangrepp, eller bör du avstå?

Förväntat antal förlorade människoliv, om du köper Bostroms siffror, blir...

- 10^8 om du attackerar,

Ett extremt otrevligt tankeexperiment

Antag att du är USA:s president, och att CIA-chefen ger dig pålitlig information om att det någonstans i Tyskland gömmer sig en terrorist som arbetar med att bygga ett domedagsvapen, som avser använda detta för att förgöra mänskligheten, och som har en chans på miljonen att lyckas med planen. Bör du attackera Tyskland med fullskaligt kärnvapenangrepp, eller bör du avstå?

Förväntat antal förlorade människoliv, om du köper Bostroms siffror, blir...

- 10^8 om du attackerar,
- minst $10^{16} \cdot 10^{-6} = 10^{10}$ om du avstår.

Ett extremt otrevligt tankeexperiment

Antag att du är USA:s president, och att CIA-chefen ger dig pålitlig information om att det någonstans i Tyskland gömmer sig en terrorist som arbetar med att bygga ett domedagsvapen, som avser använda detta för att förgöra mänskligheten, och som har en chans på miljonen att lyckas med planen. Bör du attackera Tyskland med fullskaligt kärnvapenangrepp, eller bör du avstå?

Förväntat antal förlorade människoliv, om du köper Bostroms siffror, blir...

- 10^8 om du attackerar,
- minst $10^{16} \cdot 10^{-6} = 10^{10}$ om du avstår.

(Som jag förklarar i *Here Be Dragons*, s 241, så skulle jag i denna situation ändå vägra trycka på den röda knappen.)

Vi har hittills hållt oss inom det beslutsteoretiska paradigmet för **maximering av förväntad nytta**:

Vi har hittills hållt oss inom det beslutsteoretiska paradigmet för **maximering av förväntad nytta**:

Välj den handling H som ger störst värde på

$$\sum_x U(x)P(x|H)$$

Vi har hittills hållt oss inom det beslutsteoretiska paradigmet för **maximering av förväntad nytta**:

Välj den handling H som ger störst värde på

$$\sum_x U(x)P(x|H)$$

där summan går över alla utfall x, och $U(x)$ betecknar nyttan av utfallet, medan $P(x|H)$ betecknar utfallets sannolikhet givet H.

Vi har hittills hållt oss inom det beslutsteoretiska paradigmet för **maximering av förväntad nytta**:

Välj den handling H som ger störst värde på

$$\sum_x U(x)P(x|H)$$

där summan går över alla utfall x, och $U(x)$ betecknar nyttan av utfallet, medan $P(x|H)$ betecknar utfallets sannolikhet givet H.

Men behöver vi hålla oss inom den ramen? Kanske finns det något vettigt sätt att fatta beslut utan att använda nyttofunktion U och sannolikhetsfunktion P?

Under 1900-talet utvecklades vad vi kan kalla *beslutsteorins grunder*, med en serie resultat som tillsammans visar att så länge en beslutsfattare håller sig till beslutsregler som uppfyller vissa axiom som kan anses rimliga att kräva av den som gör anspråk på **rationalitet**, så kan beslutsreglerna representeras i termer av maximering av förväntad nytta.

Under 1900-talet utvecklades vad vi kan kalla *beslutsteorins grunder*, med en serie resultat som tillsammans visar att så länge en beslutsfattare håller sig till beslutsregler som uppfyller vissa axiom som kan anses rimliga att kräva av den som gör anspråk på **rationalitet**, så kan beslutsreglerna representeras i termer av maximering av förväntad nytta.

Viktiga bidrag gjordes av Frank Ramsey (1931), Bruno de Finetti (1937), John von Neumann och Oskar Morgenstern (1944) och Leonard Savage (1951). Att gå igenom denna teori faller utanför ramen för denna kurs (men se gärna Itzhak Gilboas utmärkta bok *Theory of Decision under Uncertainty* från 2009), men...

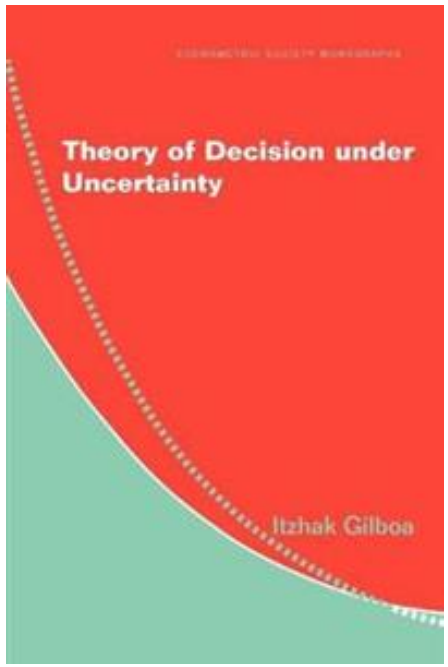
Under 1900-talet utvecklades vad vi kan kalla *beslutsteorins grunder*, med en serie resultat som tillsammans visar att så länge en beslutsfattare håller sig till beslutsregler som uppfyller vissa axiom som kan anses rimliga att kräva av den som gör anspråk på **rationalitet**, så kan beslutsreglerna representeras i termer av maximering av förväntad nytta.

Viktiga bidrag gjordes av Frank Ramsey (1931), Bruno de Finetti (1937), John von Neumann och Oskar Morgenstern (1944) och Leonard Savage (1951). Att gå igenom denna teori faller utanför ramen för denna kurs (men se gärna Itzhak Gilboas utmärkta bok *Theory of Decision under Uncertainty* från 2009), men...

...låt oss ta en titt på ett par av de centrala axiomen: **transitivitet** och **kontinuitet**.

Under 1900-talet utvecklades vad vi kan kalla *beslutsteorins grunder*, med en serie resultat som tillsammans visar att så länge en beslutsfattare håller sig till beslutsregler som uppfyller vissa axiom som kan anses rimliga att kräva av den som gör anspråk på **rationalitet**, så kan beslutsreglerna representeras i termer av maximering av förväntad nytta.

Viktiga bidrag gjordes av Frank Ramsey (1931), Bruno de Finetti (1937), John von Neumann och Oskar Morgenstern (1944) och Leonard Savage (1951). Att gå igenom denna teori faller utanför ramen för denna kurs (men se gärna Itzhak Gilboas utmärkta bok *Theory of Decision under Uncertainty* från 2009), men...



...låt oss ta en titt på ett par av de centrala axiomen: **transitivitet** och **kontinuitet**.

Transitivitet

Transitivitet

För två utfall x och y, låt

$$x \succ y$$

beteckna att beslutsfattaren föredrar utfall x framför utfall y.

Transitivitet

För två utfall x och y, låt

$$x \succ y$$

beteckna att beslutsfattaren föredrar utfall x framför utfall y.

En beslutsfattare sägs respektera **transitivitet** om det för alla utfall x, y, z, sådana att $x \succ y$ och $y \succ z$, även gäller att

$$x \succ z$$

Kontinuitet

Kontinuitet

En beslutsfattare sägs respektera **kontinuitet** om det för alla utfall x, y, z sådana att om $x \succ y \succ z$ så gäller för varje tillräckligt litet $\varepsilon > 0$ att

$$\left\{ \begin{array}{ll} x & \text{med sannolikhet } 1-\varepsilon \\ z & \text{med sannolikhet } \varepsilon \end{array} \right\} \succ y$$

samt

Kontinuitet

En beslutsfattare sägs respektera **kontinuitet** om det för alla utfall x, y, z sådana att om $x \succ y \succ z$ så gäller för varje tillräckligt litet $\varepsilon > 0$ att

$$\left\{ \begin{array}{ll} x & \text{med sannolikhet } 1-\varepsilon \\ z & \text{med sannolikhet } \varepsilon \end{array} \right\} \succ y$$

samt att

$$\left\{ \begin{array}{ll} z & \text{med sannolikhet } 1-\varepsilon \\ x & \text{med sannolikhet } \varepsilon \end{array} \right\} \prec y$$

Det kan vara frestande att invända mot kontinuitetsaxiomet med exempel som

$$\begin{aligned} x &= \text{ 1000 kr i plånboken} \\ y &= \text{ 990 kr i plånboken} \\ z &= \text{ döden.} \end{aligned}$$

Det kan vara frestande att invända mot kontinuitetsaxiomet med exempel som

$$\begin{aligned} x &= \text{ 1000 kr i plånboken} \\ y &= \text{ 990 kr i plånboken} \\ z &= \text{ döden.} \end{aligned}$$

Att $x \succ y \succ z$ kan vi vara överens om,

Det kan vara frestande att invända mot kontinuitetsaxiomet med exempel som

$$\begin{aligned} x = &\ \text{1000 kr i plånboken} \\ y = &\ \text{990 kr i plånboken} \\ z = &\ \text{döden.} \end{aligned}$$

Att $x \succ y \succ z$ kan vi vara överens om, men frågan är om det verkligen finns något tillräckligt litet $\varepsilon > 0$ så att

$$\left\{ \begin{array}{ll} x & \text{med sannolikhet } 1-\varepsilon \\ z & \text{med sannolikhet } \varepsilon \end{array} \right\} \succ y$$

Ingen skulle väl riskera att dö för ynka 10 kronor?

Det kan vara frestande att invända mot kontinuitetsaxiomet med exempel som

$$\begin{aligned} x = &\ \text{1000 kr i plånboken} \\ y = &\ \text{990 kr i plånboken} \\ z = &\ \text{döden.} \end{aligned}$$

Att $x \succ y \succ z$ kan vi vara överens om, men frågan är om det verkligen finns något tillräckligt litet $\varepsilon > 0$ så att

$$\left\{ \begin{array}{ll} x & \text{med sannolikhet } 1-\varepsilon \\ z & \text{med sannolikhet } \varepsilon \end{array} \right\} \succ y$$

Ingen skulle väl riskera att dö för ynka 10 kronor?

Men faktum är att de flesta av oss tar den sortens risker dagligen.

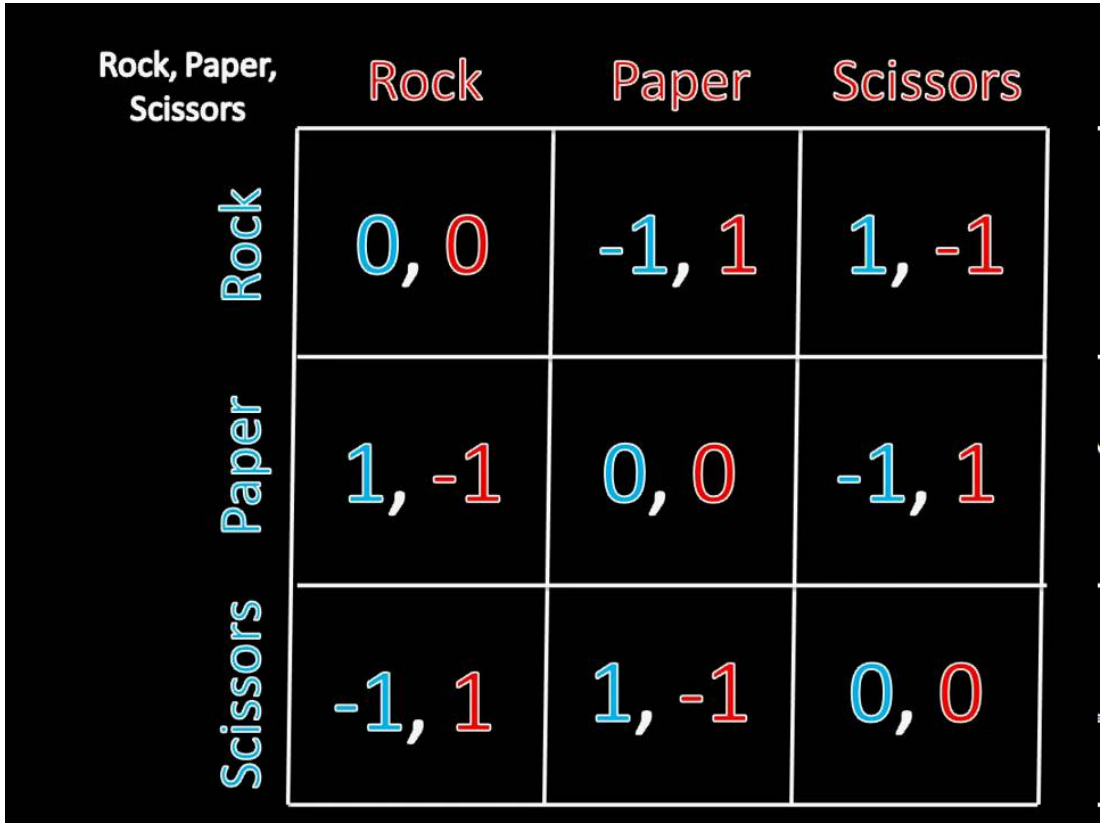
Beslutsteori vs spelteori

Beslutsteori vs spelteori

En begränsning i beslutsteorin är grundantagandet att det bara finns en enda beslutsfattare. Antar vi att det finns mer än en beslutsfattare hamnar vi i den förgrening av beslutsteorin som kallas **spelteori**.

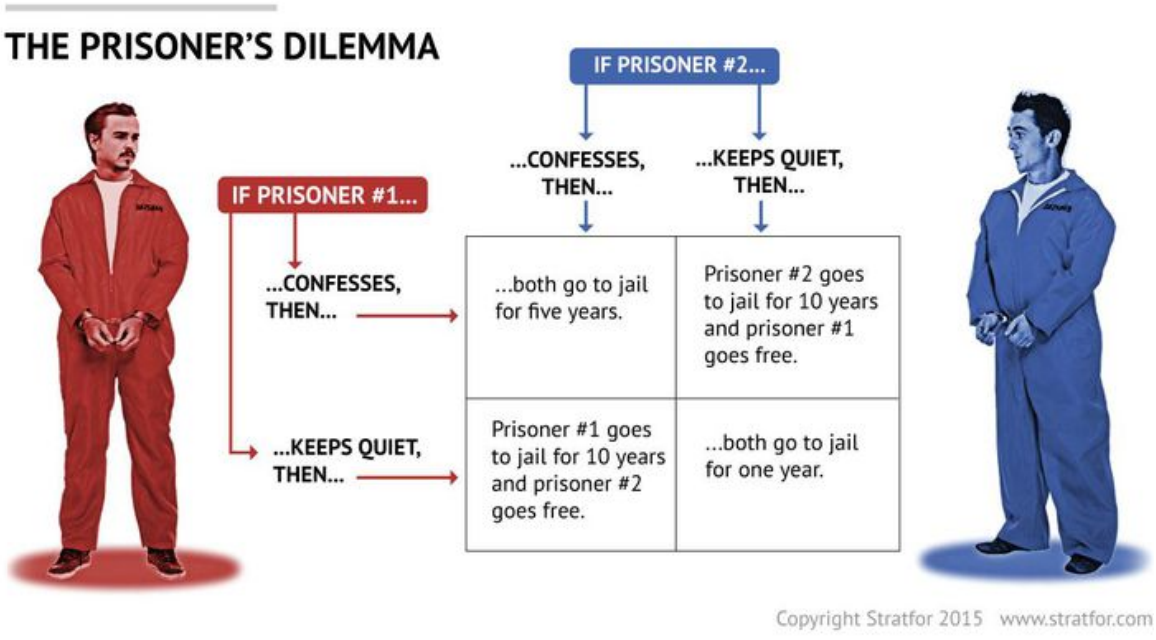
Sten-sax-påse

Sten-sax-påse



Fångarnas dilemma

Fångarnas dilemma



Fångarnas dilemma

Fångarnas dilemma

Prisoner 1 \ Prisoner 2	Cooperate	Defect
Cooperate	3, 3	0, 5
Defect	5, 0	1, 1

Nollsummespel vs icke-nollsummespel

Nollsummespel vs icke-nollsummespel

- Sten-sax-påse är ett exempel på **nollsummespel**. I sådana är spelarna renodlade antagonister.

Nollsummespel vs icke-nollsummespel

- Sten-sax-påse är ett exempel på **nollsummespel**. I sådana är spelarna renodlade antagonister.
- Men det finns också **icke-nollsummespel**, som exempelvis fångarnas dilemma, där frågor om samarbete dyker upp.

Beslutsteorins ännu djupare(?) grund

Beslutsteorins ännu djupare(?) grund

Klassisk beslutsteori säger att rationellt agerande är att, givet bakgrundsinformation b, välja det handlingsalternativ H som maximerar

$$\mathrm{E}[U|H \hookrightarrow b]\,.$$

Beslutsteorins ännu djupare(?) grund

Klassisk beslutsteori säger att rationellt agerande är att, givet bakgrundsinformation b, välja det handlingsalternativ H som maximerar

$$\mathrm{E}[U|H \hookrightarrow b]\,.$$

Men vad är $\hookrightarrow$?

Beslutsteorins ännu djupare(?) grund

Klassisk beslutsteori säger att rationellt agerande är att, givet bakgrundsinformation b, välja det handlingsalternativ H som maximerar

$$\mathrm{E}[U|H \hookrightarrow b]\,.$$

Men vad är $\hookrightarrow$?

Enklast är att låta ovanstående betyda

$$\mathrm{E}[U|H, b]\,,$$

vilket betyder att den egna handlingen H behandlas som ännu ett stycke bakgrundsinformation.

Beslutsteorins ännu djupare(?) grund

Klassisk beslutsteori säger att rationellt agerande är att, givet bakgrundsinformation b, välja det handlingsalternativ H som maximerar

$$\mathrm{E}[U|H \hookrightarrow b]\,.$$

Men vad är $\hookrightarrow$?

Enklast är att låta ovanstående betyda

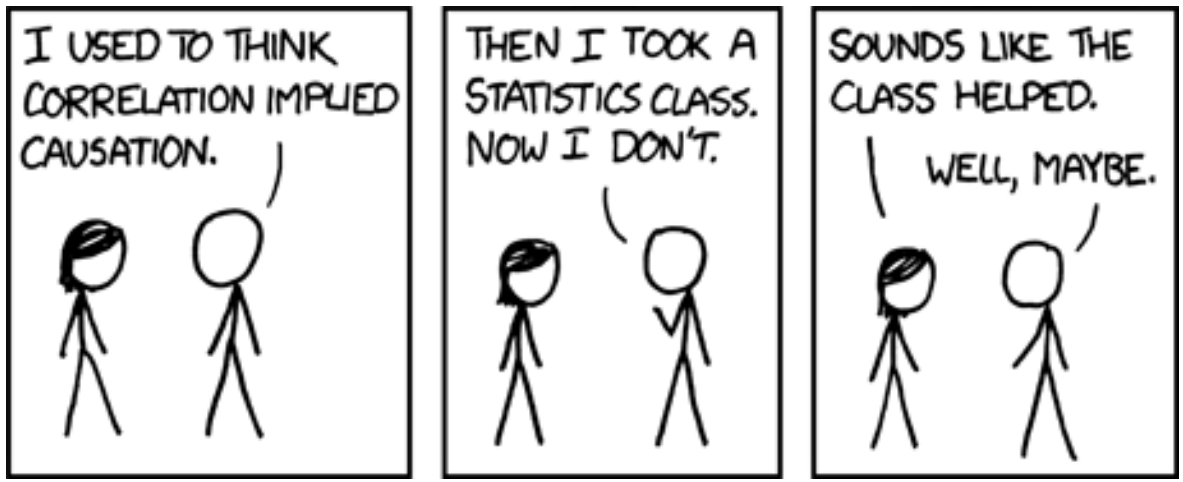
$$\mathrm{E}[U|H, b]\,,$$

vilket betyder att den egna handlingen H behandlas som ännu ett stycke bakgrundsinformation.

Förespråkare för **evidentiell beslutsteori (EDT)** gillar detta, medan förespråkare för **kausal beslutsteori (CDT)** menar att det är fel, och att valet av handling har betydelse endast via en handlings *kausala* följder.

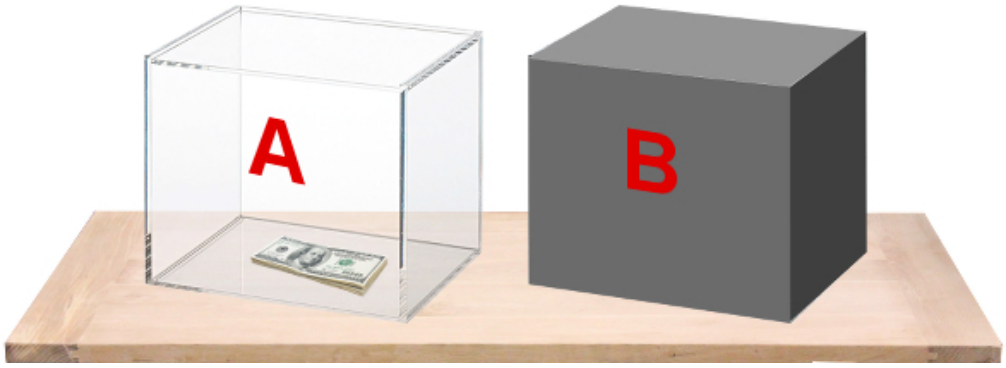
Korrelation vs kausalitet

Korrelation vs kausalitet

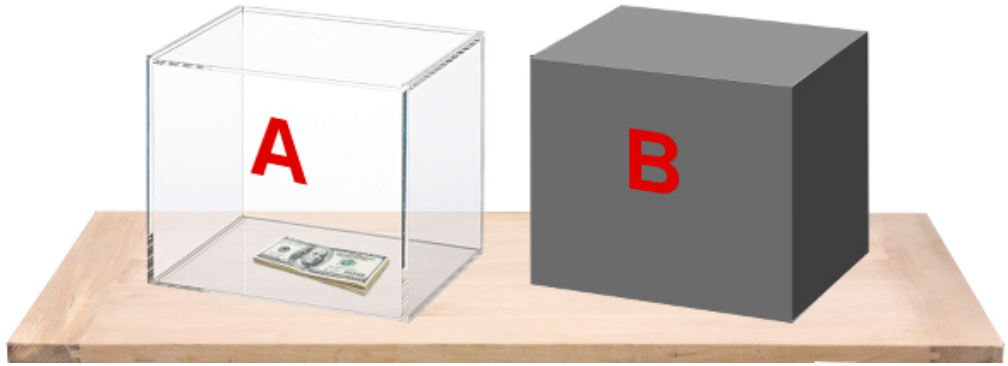


Newcombs problem

Newcombs problem

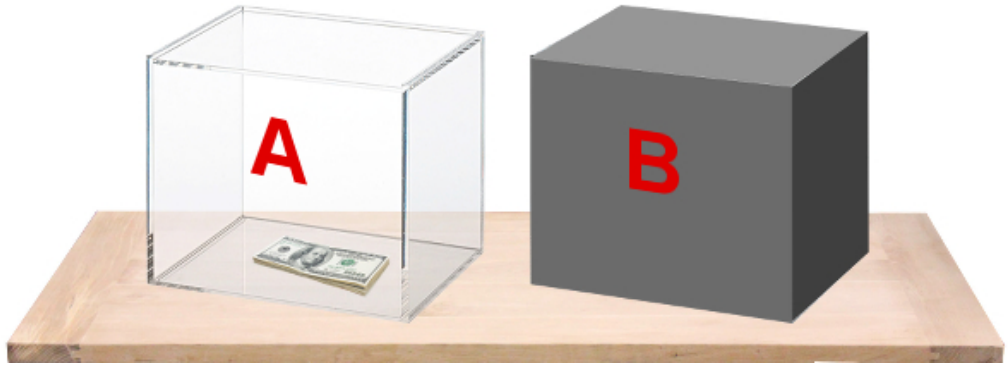


Newcombs problem



Box A är genomskinlig, och innehåller \$1000. Box B is ogenomskinlig, och innehåller antingen \$1 000 000 eller ingenting. Du kan välja att ta båda boxarna, eller enbart box B.

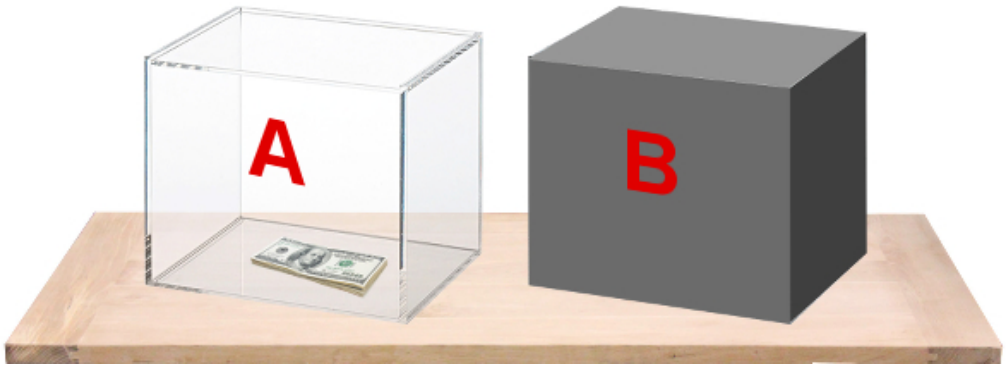
Newcombs problem



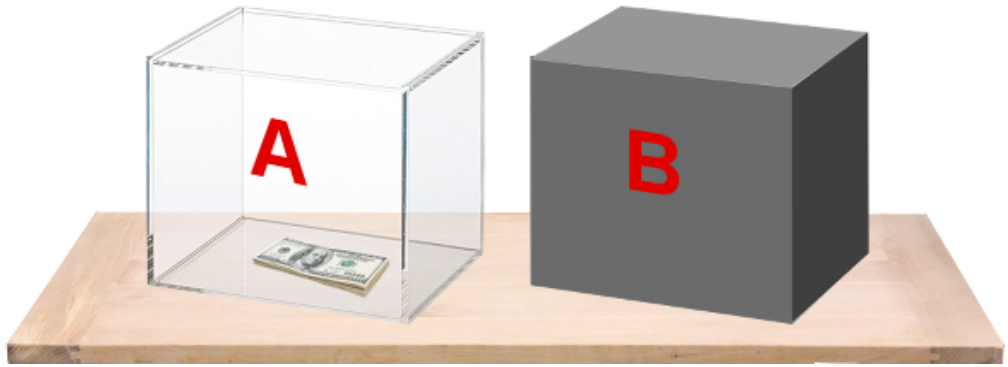
Box A är genomskinlig, och innehåller $1000. Box B is ogenomskinlig, och innehåller antingen $1 000 000 eller ingenting. Du kan välja att ta båda boxarna, eller enbart box B.

Här är haken: en övermänskligt intelligent prediktor har predikterat vad du skall välja, och lagt $1 000 000 i box B om och endast om den predikterat att du skall välja enbart box B.

Newcombs problem



Newcombs problem

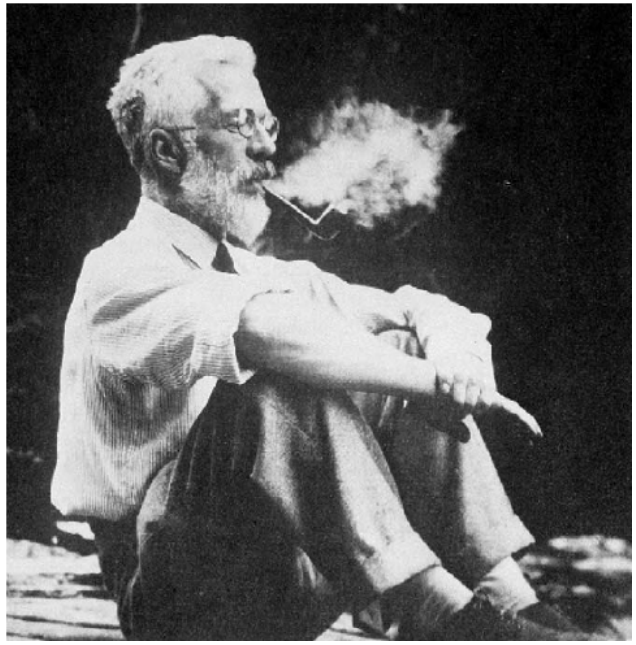


	en-boxing predikterad	två-boxing predikterad
en-boxa	\$1 000 000	\$0
två-boxa	\$1 001 000	\$1000

Robert Nozick (1969) om Newcombs problem:

"To almost everyone, it is perfectly clear and obvious what should be done. The difficulty is that these people seem to divide almost evenly on the problem, with large numbers thinking that the opposing half is just being silly."

Rökningsgenen



	bra gen	dålig gen
avstå	\$1 000 000	\$0
röka	\$1 001 000	\$1000

Fångarnas dilemma med en tvilling

	tvilling samarbetar	tvilling sviker
samarbete	\$1 000 000	\$0
svek	\$1 001 000	\$1000

EDT rekommenderar handlingar som *korrelerar med* bra utfall,

EDT rekommenderar handlingar som *korrelerar med* bra utfall, vilket betyder

- en-boxa in Newcombs problem,

EDT rekommenderar handlingar som *korrelerar med* bra utfall, vilket betyder

- en-boxa in Newcombs problem,
- avstå från rökning i Rökningsgenen,

EDT rekommenderar handlingar som *korrelerar med* bra utfall, vilket betyder

- en-boxa in Newcombs problem,
- avstå från rökning i Rökningsgenen,
- samarbeta i Fångarnas dilemma med en tvilling.

EDT rekommenderar handlingar som *korrelerar med* bra utfall, vilket betyder

- en-boxa in Newcombs problem,
- avstå från rökning i Rökningsgenen,
- samarbeta i Fångarnas dilemma med en tvilling.

CDT rekommenderar handlingar som *orsakar* bra utfall,

EDT rekommenderar handlingar som *korrelerar med* bra utfall, vilket betyder

- en-boxa in Newcombs problem,
- avstå från rökning i Rökningsgenen,
- samarbeta i Fångarnas dilemma med en tvilling.

CDT rekommenderar handlingar som *orsakar* bra utfall, vilket betyder

- två-boxa in Newcombs problem,

EDT rekommenderar handlingar som *korrelerar med* bra utfall, vilket betyder

- en-boxa in Newcombs problem,
- avstå från rökning i Rökningsgenen,
- samarbeta i Fångarnas dilemma med en tvilling.

CDT rekommenderar handlingar som *orsakar* bra utfall, vilket betyder

- två-boxa in Newcombs problem,
- röka i Rökningsgenen,

EDT rekommenderar handlingar som *korrelerar med* bra utfall, vilket betyder

- en-boxa in Newcombs problem,
- avstå från rökning i Rökningsgenen,
- samarbeta i Fångarnas dilemma med en tvilling.

CDT rekommenderar handlingar som *orsakar* bra utfall, vilket betyder

- två-boxa in Newcombs problem,
- röka i Rökningsgenen,
- svika i Fångarnas dilemma med en tvilling.

Eliezer Yudkowsky och Nate Soares (2017) föreslår ett tredje alternativ: **funktionell beslutsteori (FDT)**, som rekommenderar handlingar som du önskar algoritmen-som-är-du skulle leda till, givet målet att få så mycket nytta som möjligt under din livstid.

Eliezer Yudkowsky och Nate Soares (2017) föreslår ett tredje alternativ: **funktionell beslutsteori (FDT)**, som rekommenderar handlingar som du önskar algoritmen-som-är-du skulle leda till, givet målet att få så mycket nytta som möjligt under din livstid.

FDT är ense med EDT i Newcomb's problem och Fångarnas dilemma med en tvilling, men med CDT i Rökningsgenen.

Eliezer Yudkowsky och Nate Soares (2017) föreslår ett tredje alternativ: **funktionell beslutsteori (FDT)**, som rekommenderar handlingar som du önskar algoritmen-som-är-du skulle leda till, givet målet att få så mycket nytta som möjligt under din livstid.

FDT är ense med EDT i Newcomb's problem och Fångarnas dilemma med en tvilling, men med CDT i Rökningsgenen.

Dessutom avviker FDT från *både* EDT och CDT i diverse utpressningssituationer.

Förvirrande?

Förvirrande?

Det är helt normalt.

Nu till något helt annat!

Nu till något helt annat!

Beslutsteori beskriver *idealt rationellt handlande*. Stämmer det på människan?

Nu till något helt annat!

Beslutsteori beskriver *idealt rationellt handlande*. Stämmer det på människan?

Svar nej!

Vi mäanniskor har långtgående kognitiva förmågor – vi är bra på att tänka!

Vi mäanniskor har långtgående kognitiva förmågor – vi är bra på att tänka!

Att evolutionen skulle gynna dessa förmågor är inte svårt att föreställa sig. Evolutionen gynnar den som är bra på att hitta mat, att undvika farliga rovdjur, och att hitta (samt imponera på) någon att para sig med. Alla dessa saker kan väntas bli lättare om man är bra på att läsa av sin omvärld och dra korrekta slutsatser.

Vi mäanniskor har långtgående kognitiva förmågor – vi är bra på att tänka!

Att evolutionen skulle gynna dessa förmågor är inte svårt att föreställa sig. Evolutionen gynnar den som är bra på att hitta mat, att undvika farliga rovdjur, och att hitta (samt imponera på) någon att para sig med. Alla dessa saker kan väntas bli lättare om man är bra på att läsa av sin omvärld och dra korrekta slutsatser.

Å andra sidan...

Vi mäanniskor har långtgående kognitiva förmågor – vi är bra på att tänka!

Att evolutionen skulle gynna dessa förmågor är inte svårt att föreställa sig. Evolutionen gynnar den som är bra på att hitta mat, att undvika farliga rovdjur, och att hitta (samt imponera på) någon att para sig med. Alla dessa saker kan väntas bli lättare om man är bra på att läsa av sin omvärld och dra korrekta slutsatser.

Å andra sidan...

- Evolutionen är långt ifrån någon perfekt optimeringsalgoritm.

Vi mäanniskor har långtgående kognitiva förmågor – vi är bra på att tänka!

Att evolutionen skulle gynna dessa förmågor är inte svårt att föreställa sig. Evolutionen gynnar den som är bra på att hitta mat, att undvika farliga rovdjur, och att hitta (samt imponera på) någon att para sig med. Alla dessa saker kan väntas bli lättare om man är bra på att läsa av sin omvärld och dra korrekta slutsatser.

Å andra sidan...

- Evolutionen är långt ifrån någon perfekt optimeringsalgoritm.
- Den miljö vi evolutionärt formats för är väldigt annorlunda jämfört med dagens miljö.

Vi mäanniskor har långtgående kognitiva förmågor – vi är bra på att tänka!

Att evolutionen skulle gynna dessa förmågor är inte svårt att föreställa sig. Evolutionen gynnar den som är bra på att hitta mat, att undvika farliga rovdjur, och att hitta (samt imponera på) någon att para sig med. Alla dessa saker kan väntas bli lättare om man är bra på att läsa av sin omvärld och dra korrekta slutsatser.

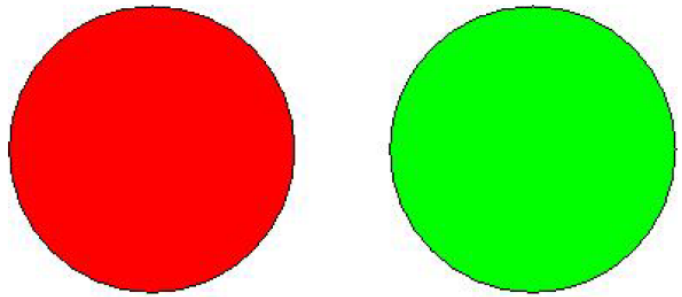
Å andra sidan...

- Evolutionen är långt ifrån någon perfekt optimeringsalgoritm.
- Den miljö vi evolutionärt formats för är väldigt annorlunda jämfört med dagens miljö.

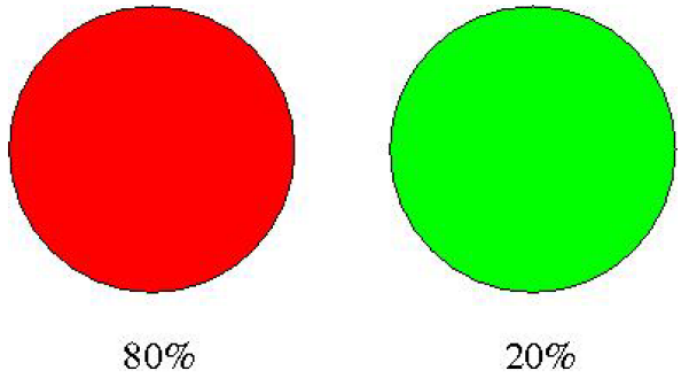
Därför är det inte så konstigt om våra hjärnor har det jag (lite provokativt) kallar **felkalibreringar**.

Ett känt exempel:
alltför “trigger-happy” mönsterdetektering

Ett känt exempel: alltför “trigger-happy” mönsterdetektering

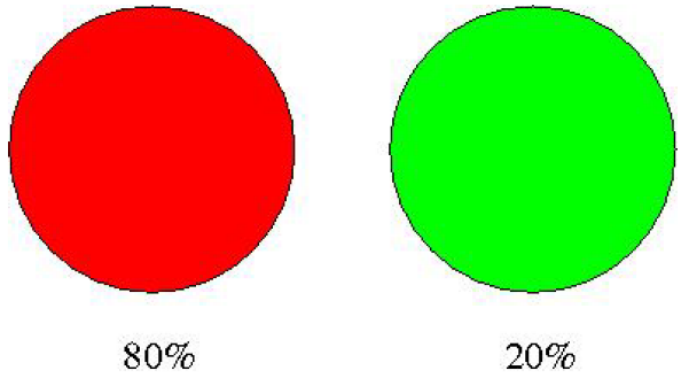


Ett känt exempel: alltför “trigger-happy” mönsterdetektering



Råttor och duvor: Identifierar snabbt vilken lampa som lyser oftast, och gissar därefter rätt 80% av gångerna.

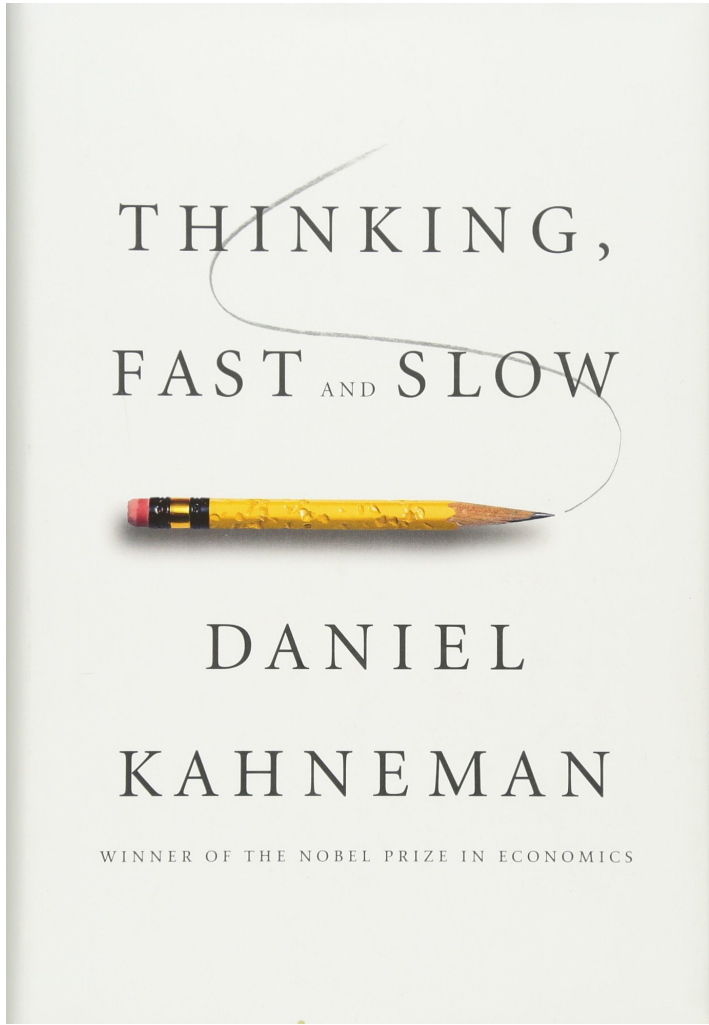
Ett känt exempel: alltför "trigger-happy" mönsterdetektering



Råttor och duvor: Identifierar snabbt vilken lampa som lyser oftast, och gissar därefter rätt 80% av gångerna.

Människor: Söker (typiskt) efter mönster, identifierar proportionen av röd respektive grön, och gissar därefter rätt $0.8 \cdot 0.8 + 0.2 \cdot 0.2 = 0.64 + 0.04 = 0.68 = 68\%$ av gångerna.

Den främsta källan när det gäller den mänskliga hjärnans felkalibreringar är denna:



När det gäller exempel med särskild relevans för denna kurs vill jag dock ännu starkare rekommendera Eliezer Yudkowskys uppsats **Cognitive biases potentially affecting judgement of global risks** från 2008.

När det gäller exempel med särskild relevans för denna kurs vill jag dock ännu starkare rekommendera Eliezer Yudkowskys uppsats **Cognitive biases potentially affecting judgement of global risks** från 2008. Exempel som tas upp där inkluderar bland annat

- *overconfidence* (övertro på det egna kunskapsläget)

När det gäller exempel med särskild relevans för denna kurs vill jag dock ännu starkare rekommendera Eliezer Yudkowskys uppsats **Cognitive biases potentially affecting judgement of global risks** från 2008. Exempel som tas upp där inkluderar bland annat

- *overconfidence* (övertro på det egna kunskapsläget)
- *availability bias* (tillgänglighetsbias)

När det gäller exempel med särskild relevans för denna kurs vill jag dock ännu starkare rekommendera Eliezer Yudkowskys uppsats **Cognitive biases potentially affecting judgement of global risks** från 2008. Exempel som tas upp där inkluderar bland annat

- *overconfidence* (övertro på det egna kunskapsläget)
- *availability bias* (tillgänglighetsbias)
- tidsoptimism