MVE420:
Nya teknologier, global risk och mänsklighetens framtid

http://www.math.chalmers.se/Math/Grundutb/CTH/mve420/1819/

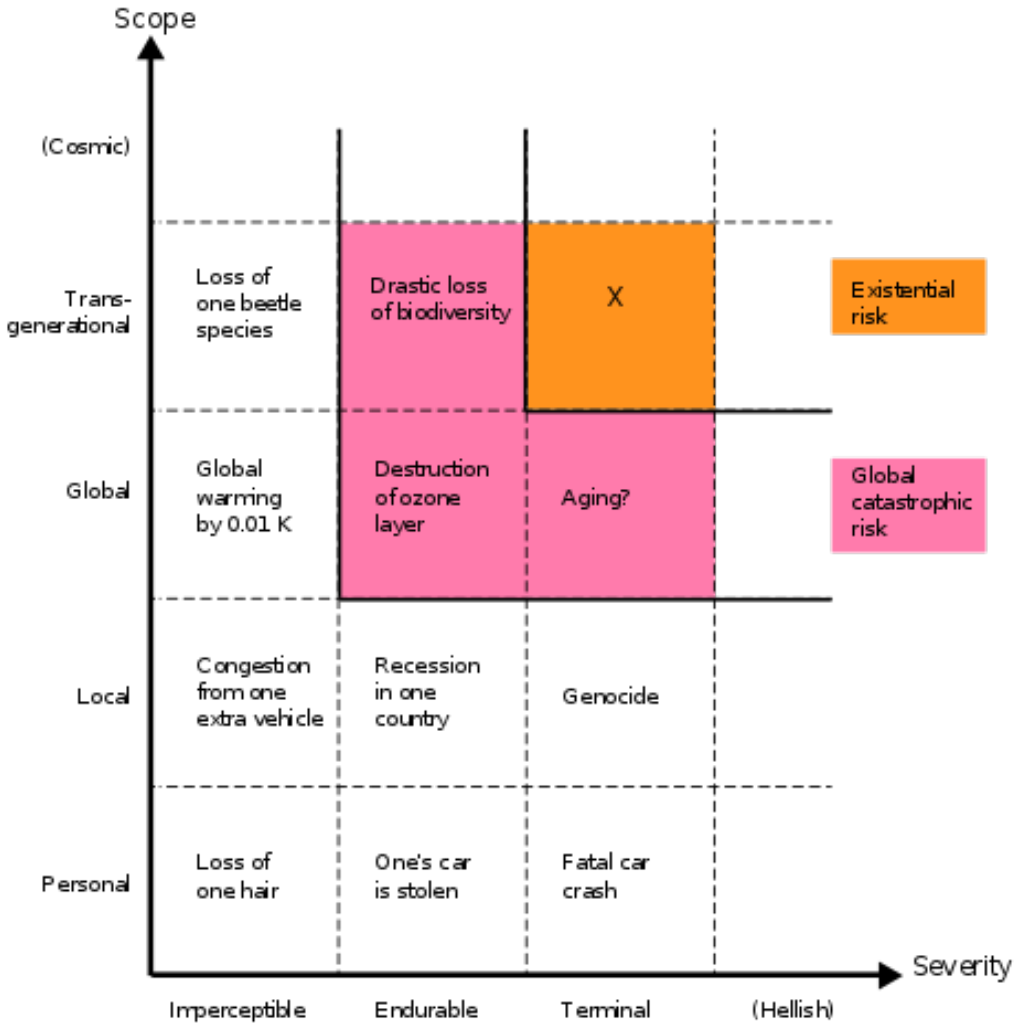
Föreläsning om

Besultsteori, rationalitet och existentiell risk

3 maj 2019

Olle Häggström

Bostroms (2013) schematiska klassificering av risker:



Bostroms definition av existentiell risk:

> *An existential risk is one that threatens the premature extinction of Earth-originating intelligent life, or the permanent and drastic destruction of its potential for desirable future development.*

Bostroms definition av existentiell risk:

> *An existential risk is one that threatens the premature extinction of Earth-originating intelligent life, or the permanent and drastic destruction of its potential for desirable future development.*

Några saker att fundera kring:

Bostroms definition av existentiell risk:

> *An existential risk is one that threatens the premature extinction of Earth-originating intelligent life, or the permanent and drastic destruction of its potential for desirable future development.*

Några saker att fundera kring:

- Varför *"Earth-originating intelligent life"* snarare än helt enkelt människor?

Bostroms definition av existentiell risk:

> *An existential risk is one that threatens the premature extinction of Earth-originating intelligent life, or the permanent and drastic destruction of its potential for desirable future development.*

Några saker att fundera kring:

- Varför *"Earth-originating intelligent life"* snarare än helt enkelt människor?
- Vad för slags händelser (annat än utrotning) kan *"permanent and drastic destruction of its potential"* syfta på?

Bostroms definition av existentiell risk:

> *An existential risk is one that threatens the premature extinction of Earth-originating intelligent life, or the permanent and drastic destruction of its potential for desirable future development.*

Några saker att fundera kring:

- Varför *"Earth-originating intelligent life"* snarare än helt enkelt människor?
- Vad för slags händsleser (annat än utrotning) kan *"permanent and drastic destruction of its potential"* syfta på?
- Varför *"premature"*?

Bostroms definition av existentiell risk:

> *An existential risk is one that threatens the premature extinction of Earth-originating intelligent life, or the permanent and drastic destruction of its potential for desirable future development.*

Några saker att fundera kring:

- Varför *"Earth-originating intelligent life"* snarare än helt enkelt människor?
- Vad för slags händsleser (annat än utrotning) kan *"permanent and drastic destruction of its potential"* syfta på?
- Varför *"premature"*?
- Vad menas med *"desirable"*?

En möjlig kritik av Bostroms klassificering och definition är att de inte tar hänsyn till hur sannolika eller akuta de olika riskerna är.

En möjlig kritik av Bostroms klassificering och definition är att de inte tar hänsyn till hur sannolika eller akuta de olika riskerna är.

Alexey Turchin och David Denkenberger tar sig an detta i uppsatsen *Global catastrophic and existential risks communications scale* (2018).

Colour coding	**Explanation of colors**	**Classification of prevention actions** (by Tonn and Steifel's)	**Indicative risk**	**Probability interval** for human extinction risks in next 100 years, if known
Purple	**Immediate extreme risks** of human extinction	Extreme war footing, economy organized around reducing human extinction risk	Imminent risk of WW3 beginning (like Cuban missile crisis);	Imminent
Red	**High risks** of existential catstophe	Rationing, population control, major command and control regulations	AI control problem	10-100%
Orange	**Serious risks** of global catastrophe, large prevention efforts	Manhattan scale projects	Full scale nuclear war producing human extinction	0.1-10% (once in ~10 000 years event)
Yellow	**Medium risks**, which require some prevention activity	Major programs (e.g., carbon tax) and major public investments	Supervolcanic eruption (Toba size event) of extinction level size	0.001-0.1% (once in ~1 million years event)
Green	**Small risks**, which require observation	Minor tax incentives, deployment programs;	Asteroid danger (Chicxulub-size event)	0.00001-0.001% (once in ~100 million years event)
White	**Theoretical risks**	Do nothing;	Sun becomes red giant	Less than 0.00001% (once in >1 billion years event)

Table 1. X risks communication scale: explanation of colors and coordination with Tonn and Steifel's classification of prevention actions

Size of the catastrophe: / *Probability interval in the next 100 years (equivalent timing):*	All possible civilizations in the visible Universe destroyed (including humanity)	All life on Earth destroyed (including humanity)	Human extinction	Civilization collapse (small group of people survives)	Global catastrophe (above 1 billion victims)
This row represents immediate near-term danger					
0.1 - 1					10% global agricultural shortfall
0.01 - 0.1		**Non-aligned AI**	Synthetic biology risks	**Full scale nuclear war**	Natural pandemic
0.001 - 0.01 (100k-10k years)		Nanotech			**Global warming**
10E-4 - 0.001 (1 mln-100k years)				Super volcanic eruption	
10E-5 - 10E-4 (10 mln-1 mln years)				**Asteroid impact**	
10E-6 - 10E-5 (100 mln-10 mln years)					
10E-7 - 10E-6 (1 bn - 100 million years)					
10E-8 - 10E-7 (10 bn- 1bn years)		Collider black hole or strangelet			
10E-8 (around 10 billion years)	Collider false vacuum decay	**Sun becomes red giant**			

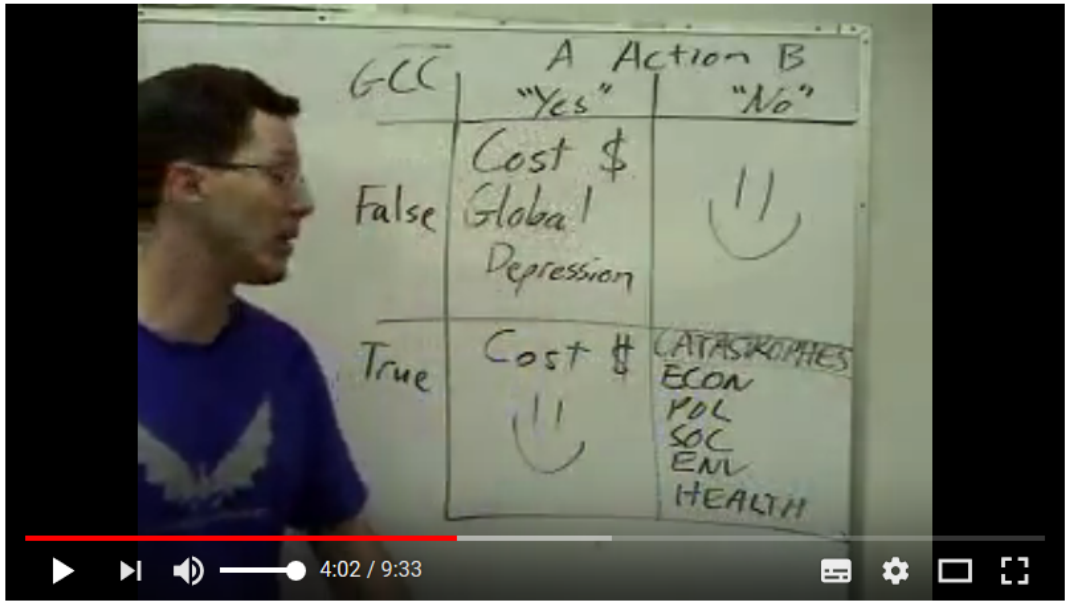
En första övning i beslutsteori

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

Sök



The Most Terrifying Video You'll Ever See

7 013 746 visningar

24 TN 6,9 TN DELA

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

Om värld I har sannolikhet p och värld II har sannolikhet $1-p$ så 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

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

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 (och oftast konkav) nyttofunktion är ganska vanligt, och vi träffade på det redan i min föreläsning om klimatförändringar och samarbete (Ramseys formel $r = \eta g + \delta$).

Blaise Pascal (1623-1662)

Pascal's Wager

Pascal's Wager

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

$$\begin{array}{lll} \textbf{tro} & \text{förväntat utbyte} & +\infty \cdot \varepsilon - 1(1-\varepsilon) = +\infty \\ \textbf{otro} & \text{förväntat utbyte} & -\infty \cdot \varepsilon + 1(1-\varepsilon) = -\infty \end{array}$$

...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 och framtida generationer) för antalet framtida människoliv om vi inte går under –

10^{17} med "business as usual", 10^{34} med rymdkolonisering, och 10^{54} med uppladdning

– så hamnar man lätt i just sådana.

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...

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 $10^{17} \cdot 10^{-6} = 10^{11}$ 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 $10^{17} \cdot 10^{-6} = 10^{11}$ 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**: 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

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

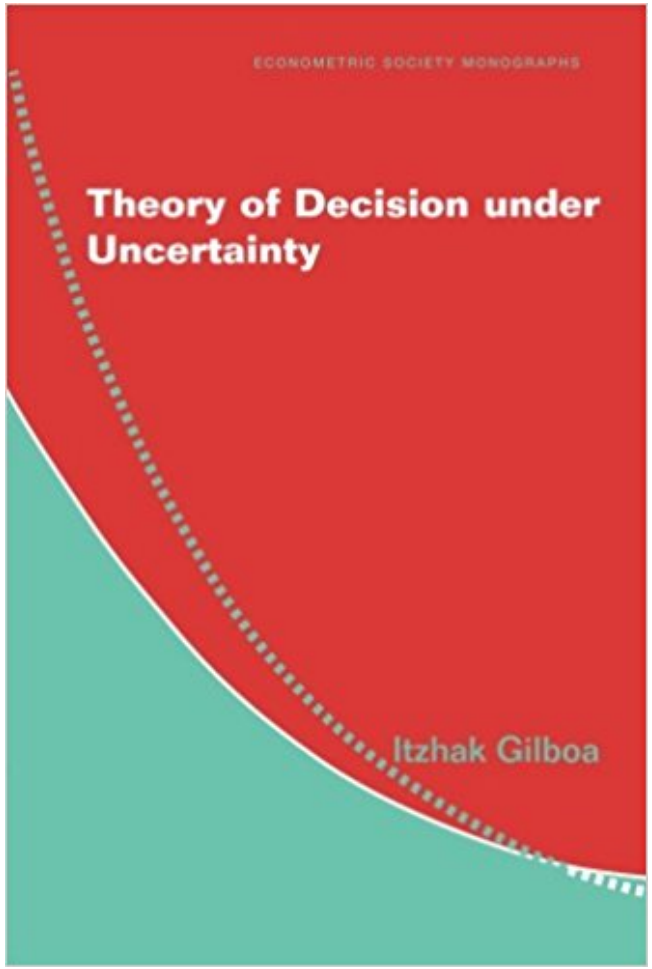
där summan går över alla utfall x, och $U(x)$ betecknar nyttan av utfallet, medan $P(x|H)$ betecknar utfallets sannolikhet givet H.

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.

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**.


ECONOMETRIC SOCIETY MONOGRAPHS
Theory of Decision under Uncertainty
Itzhak Gilboa

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

$$x \succ z$$

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

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

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

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

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

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

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

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

Det kan vara frestande att invända mot kontinuitetsaxiomet med exempel som

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

ty ingen skulle väl riskera att dö för ynka 10 kronor?

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

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

Det kan vara frestande att invända mot kontinuitetsaxiomet med exempel som

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

ty 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

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**.

WHAT A PAYOFF MATRIX LOOKS LIKE

Player 1

		Player 2		
		Rock	Paper	Scissors
Rock		0 0	1 -1	-1 1
Paper		-1 1	0 0	1 -1
Scissors		1 -1	-1 1	0 0

Nollsummespel vs icke-nollsummespel

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