MVE420:
Nya teknologier, global risk och mänsklighetens framtid

http://www.math.chalmers.se/Math/Grundutb/CTH/mve420/1415/

Föreläsning om

Felkalibreringar i den mänskliga hjärnan som försvårar riskbedömning

21 april 2015

Olle Häggström

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

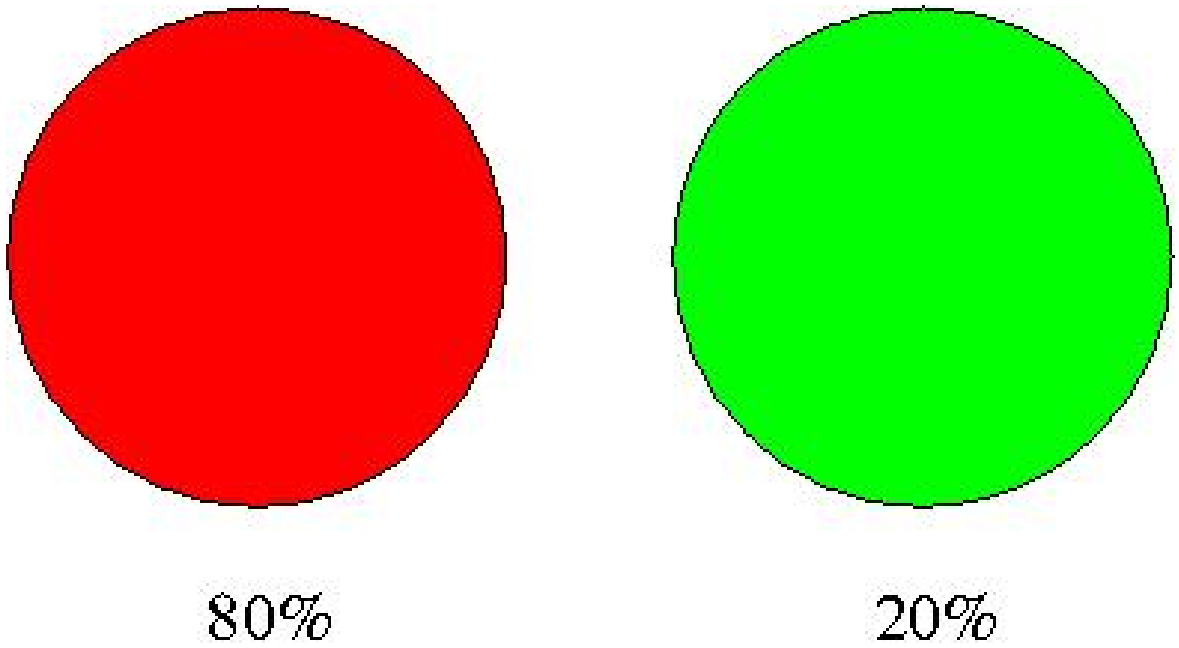
Att evolutionen skulle gynna dessa förmågor är inte svårt att föreställa sig. Evolutionen gynnar den som är bra på att hitta mat, att undvika farliga rovdjur, och att hitta (samt imponera på) någon att para sig med. Alla dessa saker kan väntas bli lättare om man är bra på att läsa av sin omvärld och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

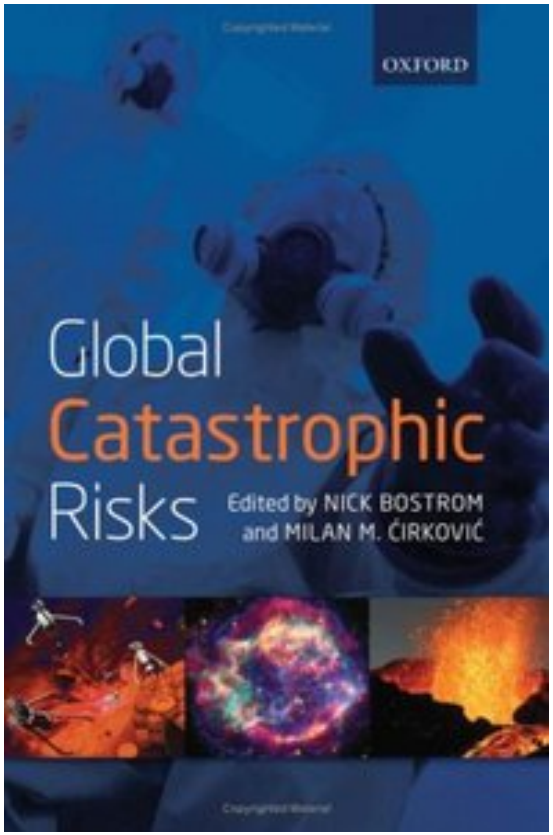
Ett första exempel: alltför "trigger-happy" mönsterdetektering.



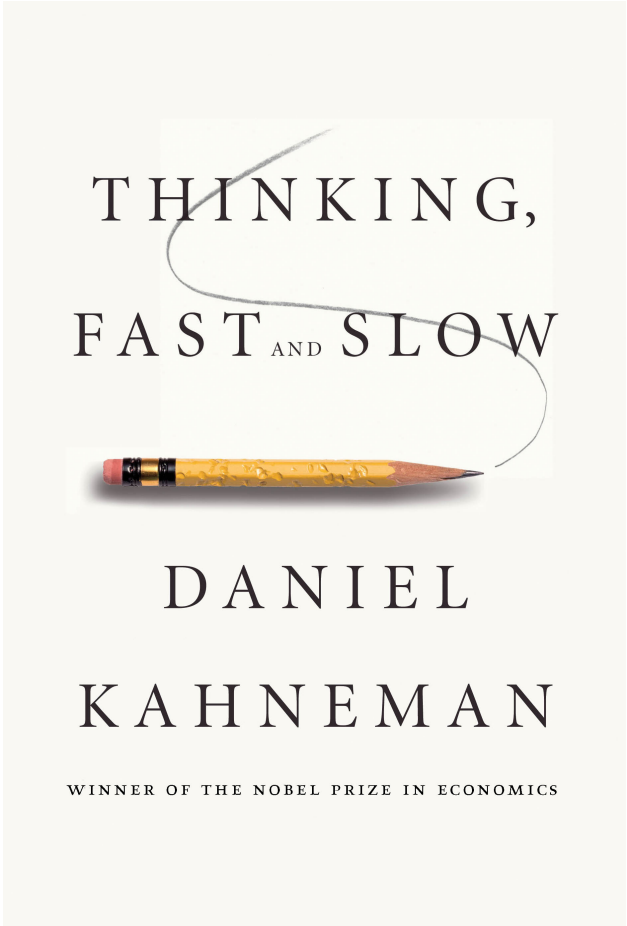
Råttor och duvor: Identifierar snabbt vilken lampa som lyser oftast, och gissar därefter rätt 80% av gångerna.

Människor: Söker (typiskt) efter mönster, identifierar proportionen av röd espektive grön, och gissar därefter rätt $0{,}8 \cdot 0{,}8 + 0{,}2 \cdot 0{,}2 = 0{,}64 + 0{,}04 = 0{,}68 = 68\%$ av gångerna.

Eliezer Yudkowskys uppsats **Cognitive biases potentially affecting judgement of global risks** finns på s 91–119 i nedanstående bok från 2008, samt på `https://intelligence.org/files/CognitiveBiases.pdf`



En annan utmärkt källa på detta område, mer generell och mindre inriktad på just bedömning av globala katastrofrisker, är denna:



En inflytelserik felkalibrering är vad Wikipedia kallar **“tillgänglighetsheuristik”** (på engelska *availability bias*).

Yudkowsky ber oss ta ställning till följande:

Suppose you randomly sample a word of three or more letters from an English text. Is it more likely that the word starts with an R (“rope”), or that R is its third letter (“park”)?

Ord med R som första bokstav är lättare att komma på, och de flesta svarar därför att dessa är vanligare.

(I själva verket är ord med R som tredje bokstav vanligare.)

Ett känt fenomen är folks bristande benägenhet att försäkra sig mot översvämning, även när priset är starkt subventionerat. Detta beror troligen i hög grad på att översvämningsnivåer högre än vad som upplevts i mannaminne ses som i princip omöjliga.

Hur skall vi då vänta oss att de bedömer risker för katastrofer som leder till mänskligheten undergång...?

Nästa exempel: Nu vill jag använda er som försökskanier.

Lägg ifrån er smartphones, etc, och ta ställning till följande fråga:

Hur stort är avståndet, enligt Eniro Kartsök,
snabbaste bilvägen mellan Malmö och Sundsvall?

Jag vill veta inte bara *en* siffra, utan *två*, en övre och en undre gräns som tillsammans omsluter ett 98%-igt subjektivt trolighetsintervall.

I en stor studie från 1982 missade drygt 42% av de angivna trolighetsintervallen det verkliga värdet, jämfört med de 2% som skulle ha blivit fallet om försökspersonerna hade en välavvägd uppfattning om det egna kunskapsläget.

Det visade sig möjligt att pressa ned felfrekvensen något genom att istället för 98%-iga trolighetsintervall fråga efter 99,9%-iga. Då blev det ”bara” 40% missar.

Nästa exempel: förankringseffekt.

En grupp försökspersoner fick följande frågor om hur höga amerikanska sekvojaträd kunde bli:

> *Är det högsta trädet högre eller lägre än 366 meter?*
> *Hur högt tror du att det högsta trädet är?*

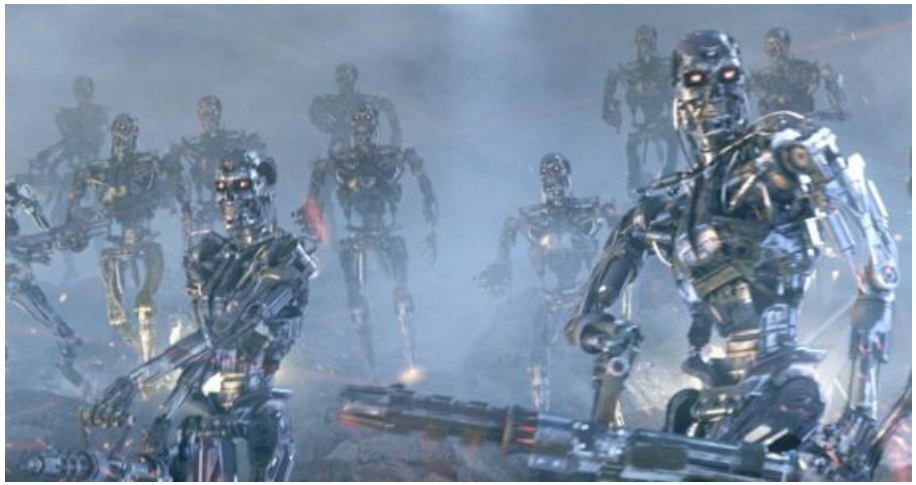
En annan grupp fick istället följande frågor:

> *Är det högsta trädet högre eller lägre än 55 meter?*
> *Hur högt tror du att det högsta trädet är?*

Den första gruppen besvarade följdfrågan med i genomsnitt 257 meter, den andra med i genomsnitt 86 meter.

Det är fullt möjligt att rationellt försvara olika gissningar i de två fallen ovan, men vad skall man då säga om Wikipedias nästa exempel?

Försökspersoner ombads först skriva ner de två sista siffrorna i sina personnummer och ta ställning till om de skulle betala detta antal dollar för föremål av obekant värde (t.ex. vin, choklad eller datorutrustning). Efter det uppmanades de att ge bud på dessa föremål. De med högre tvåsiffrigt tal i personnumret gav bud 60 till 120 procent högre än de med lägre tvåsiffrigt tal.



Denna bild verkar vara mer eller mindre obligatorisk i diskussion av risker i samband med ett AI-genombrott. Vad har det för effekt på våra bedömningar av dessa risker? Kan vi drabbas av förankringseffekt?

Nästa exempel: konjunktionsbias

För godtyckliga händelser A och B gäller sambandet

$$\mathbf{P}(A \text{ och } B) \leq \mathbf{P}(A)\,.$$

Om exempelvis $A =$ "jag är minst 190 cm lång" och $B =$ "jag väger minst 80 kg", så får vi alltså att händelsen

"jag är minst 190 cm lång"

är *minst* lika sannolik som

"jag är minst 190 cm lång och väger minst 80 kg".

Människors spontana sannolikhetsuppskattningar tenderar ofta att bryta mot sambandet $\mathbf{P}(A \text{ och } B) \leq \mathbf{P}(A)$.

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discimination and social justice, and also participated in antinuclear demonstrations.

Rank the following statements from most probable to least probable:

(1) Linda is a teacher in elementary school.

(2) Linda works in a bookstore and takes Yoga classes.

(3) Linda is active in the feminist movement.

(4) Linda is a psychiatric social worker.

(5) Linda is a member of the League of Women Voters.

(6) Linda is a bank teller.

(7) Linda is an insurance salesperson.

(8) Linda is a bank teller and is active in the feminist movement.

89% av tillfrågade försökspersoner rankade (8) som mer sannolik än (6). Det går att tolka Linda-exemplet så att ett sådant svar blir rationellt, men konjunktionsbiasfenomenet är ganska robust och har konstaterats i en mängd olika försökssituationer. Här ett annat:

> *Betrakta kast med en tärning som har fyra gröna sidor och två röda. Välj ut en av följande sekvenser; du får $25 om successiva kast med tärningen ger upphov till just den sekvensen:*
>
> *(1) RGRRR*
> *(2) GRGRRR*
> *(3) GRRRRR*

65% av försökspersonerna valde (2), trots att

$$\mathbf{P}((2)) = \mathbf{P}(\text{först G, därefter } (1)) \leq \mathbf{P}((1)).$$

Mer generellt tenderar vi att överskatta sannolikheten för händelser av typen "A_1 och A_2 och ... och A_n", och vi tenderar att underskatta sannolikheten för "A_1 eller A_2 eller ... eller A_n".

Yudkowsky ger exemplet

> *We don't need to worry about nanotechnologic war, because a UN commission will initially devlop the technology and prevent its proliferation until such time as an active shield is developed, capable of defending against all accidental and malicious outbreaks that contemporary nanotechnology is capable of producing, and this condition will persist indefinitely*

...och kommenterar att "vivid, specific scenarios can inflate our probability estimates of security, as well as misdirecting defensive investments into needlessly narrow or implausibly detailed risk scenarios".

Tidsoptimism: I en studie ombads studenter ange tidpunkter för när de kände sig 50%, 75% och 99% säkra på att de skulle bli klara med ett uppsatsarbete. 13% blev klara till sin 50%-deadline, 19% till sin 75%-deadline, och 45% till sin 99%-deadline.

Kanske kan tidsoptimism förklaras med konjunktionsbias. Om jag gör bedömningen att jag är 75% säker på att bli klar med tentarättningen före 1 juni, så bygger det möjligen på en överskattning av t.ex. $\mathbf{P}(A_1 \text{ och } A_2 \text{ och } A_3 \text{ och } A_4)$, där

$A_1 =$ "jag kommer att vara fullt frisk och arbetsför fram till 1 juni",

$A_2 =$ "jag kommer inte att drabbas av någon akut familjekris fram till 1 juni",

$A_3 =$ "jag kommer inte att falla för frestelsen att låta mig distraheras av roligare arbetsuppgifter"

$A_4 =$ "tentan kommer att visa sig precis så lätträttad som jag hoppas".

Ett sista exempel: kvantitetsblindhet.

För att mäta folks inställning i t.ex. miljöfrågor gör ekonomer ibland *villighet att betala*-studier.

I en sådan studie ställdes frågan

> *2000 flyttfåglar dör årligen genom drunkning i oskyddade oljedammar, som fåglarna uppfattar som vatten. Detta kan undvikas genom att täcka oljedammarna med nät. Vad skulle du vara villig att betala för en sådan åtgärd?*

Andra försökspersoner fick samma fråga, men med 20 000 eller 200 000 fåglar istället för 2000. Deras genomsnittliga betalningsvillighet blev

2000	\$80
20 000	\$78
200 000	\$88

Albert Szent-Győrgyi:

> *I am deeply moved if I see one man suffering and would risk my life for him. Then I talk impersonally about the possible pulverization of our big cities, with a hundred million dead. I am unable to multiply one man's suffering by a hundred million.*

Yudkowsky kommenterar:

> *Human emotions take place within an analog brain which cannot release enough neurotransmitters to feel emotion a thousand times as strong as the grief of one funeral. A prospective risk going from 10,000,000 deaths to 100,000,000 deaths does not multiply by ten the strength of our determination to stop it. It adds just one more zero on paper for our eyes to glaze over.*

Denna kvantitetsblindhet utforskas vidare i Paul Slovics läsvärda uppsats **"If I look at the mass I will never act": Psychic numbing and genocide** från 2007.