MVE420:
Nya teknologier, global risk och mänsklighetens framtid

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

Föreläsning om

Felkalibreringar i den mänskliga hjärnan som försvårar riskbedömning

9 april 2019

Olle Häggström

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

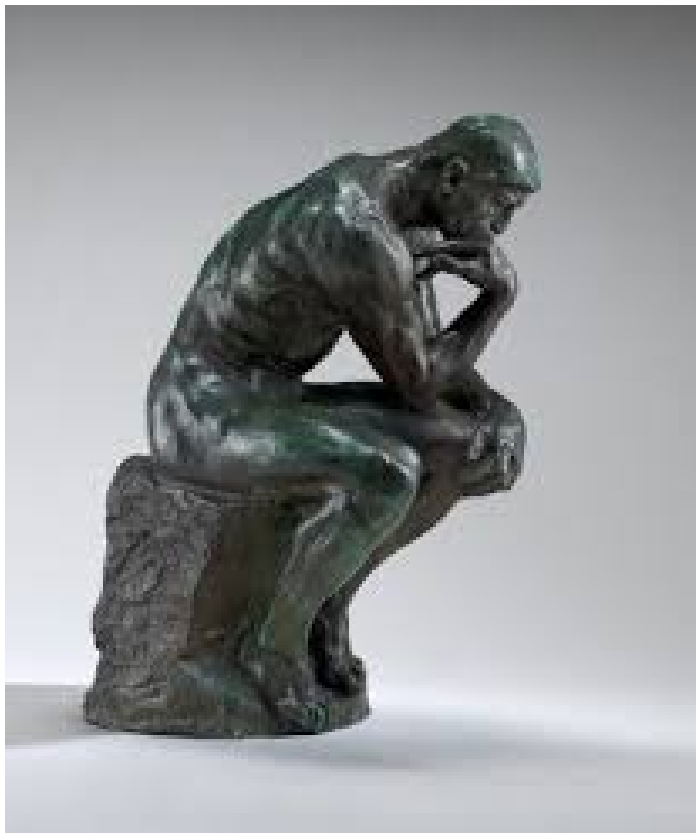
Att evolutionen skulle gynna dessa förmågor är inte svårt att föreställa sig. Evolutionen gynnar den som är bra på att hitta mat, att undvika farliga rovdjur, och att hitta (samt imponera på) någon att para sig med. Alla dessa saker kan väntas bli lättare om man är bra på att läsa av sin omvärld och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

Detta för oss in på kognitions- och beteendevetenskapen...



...som dock verkar ha drabbats hårt av den så kallade **replikerbarhetskrisen**, vilken för ett par år sedan gav upphov till många braskande rubriker.

- **Science is broken – but it can be fixed** (*Wired*, oktober 2014)
- **The statistical crisis in science** (*American Scientist*, november 2014)
- **Online debate erupts to ask: is science broken?** (*Nature*, mars 2015)
- **Is science in crisis?** (*Baltimore Sun*, juli 2015)
- **The reproducibility crisis: cognitive scientist, heal thyself** (*The Federalist*, september 2015)
- **A quick guide to the replication crisis in psychology** (*Psychology Today*, september 2015)

Analytic Thinking Promotes Religious Disbelief

Will M. Gervais* and Ara Norenzayan*

Scientific interest in the cognitive underpinnings of religious belief has grown in recent years. However, to date, little experimental research has focused on the cognitive processes that may promote religious disbelief. The present studies apply a dual-process model of cognitive processing to this problem, testing the hypothesis that analytic processing promotes religious disbelief. Individual differences in the tendency to analytically override initially flawed intuitions in reasoning were associated with increased religious disbelief. Four additional experiments provided evidence of causation, as subtle manipulations known to trigger analytic processing also encouraged religious disbelief. Combined, these studies indicate that analytic processing is one factor (presumably among several) that promotes religious disbelief. Although these findings do not speak directly to conversations about the inherent rationality, value, or truth of religious beliefs, they illuminate one cognitive factor that may influence such discussions.

Although most people fervently believe in God or gods, there are nonetheless hundreds of millions of nonbelievers worldwide (*1*), and belief and disbelief fluctuate across situations and over time (*2*). Religious belief and disbelief are likely complex, multidetermined, psychologically and culturally shaped phenomena, yet to date little experimental research has explored the specific cognitive underpinnings of religious disbelief (*3*, *4*). Here we begin to address this important gap in the literature by applying a dual-process cognitive framework, which predicts that analytic thinking strategies might be one potent source of religious disbelief.

According to dual-process theories of human thinking (*5*, *6*), there are two distinct but interacting systems for information processing. One (System 1) relies upon frugal heuristics yielding intuitive responses, while the other (System 2) relies upon deliberative analytic processing. Although both systems can at times run in parallel (*7*), System 2 often overrides the input of system 1 when analytic tendencies are activated and cognitive resources are available. Dual-process theories have been successfully applied to diverse domains and phenomena across a wide range of fields (*5*, *6*, *8*, *9*).

Available evidence and theory suggest that a converging suite of intuitive cognitive processes facilitate and support belief in supernatural agents, which is a central aspect of religious beliefs worldwide (*10*–*13*). These processes include intuitions about teleology (*14*), mind-body dualism (*13*), psychological immortality (*15*), and mind perception (*16*, *17*). Religious belief therefore bears many hallmarks of System 1 processing.

If religious belief emerges through a converging set of intuitive processes, and analytic processing can inhibit or override intuitive processing, then analytic thinking may undermine intuitive support for religious belief. Thus, a dual-process account predicts that analytic thinking may be one source of religious disbelief. Recent evidence is consistent with this hypothesis (*4*), finding that individual differences in reliance on intuitive thinking predict greater belief in God, even after controlling for relevant socio-demographic variables. However, evidence for causality remains rare (*4*). Here we report five studies that present empirical tests of this hypothesis.

We adopted three complementary strategies to test for robustness and generality. First, study 1 tested whether individual differences in the tendency to engage analytic thinking are associated with reduced religious belief. Second, studies 2 to 5 established causation by testing whether various experimental manipulations of analytic processing, induced subtly and implicitly, encourage religious disbelief. These manipulations of analytic processing included visual priming, implicit priming, and cognitive disfluency (*18*, *19*). Third, across studies, we assessed religious belief using diverse measures that focused primarily on belief in and commitment to religiously endorsed supernatural agents. Samples consisted of participants from diverse cultural and religious backgrounds (*20*).

Study 1 was a correlational study with Canadian undergraduates ($N = 179$). We correlated performance on an analytic thinking task with three related, but distinct, measures of religious belief. The analytic thinking task (*6*) contains three problems that require participants to analytically override an initial intuition. This task was designed to specifically measure analytic processing because an intuitive reading of each problem invites a quick and easy, yet incorrect, response that must be analytically overridden (Table 1). Furthermore, experimental manipulations known to induce analytic processing

University of British Columbia, Vancouver, BC V6T1Z4, Canada.

*To whom correspondence should be addressed. E-mail: will@psych.ubc.ca (W.M.G.); ara@psych.ubc.ca (A.N.)

PLOS ONE

RESEARCH ARTICLE

Direct replication of Gervais & Norenzayan (2012): No evidence that analytic thinking decreases religious belief

Clinton Sanchez[1,2], Brian Sundermeier[3], Kenneth Gray[4], Robert J. Calin-Jageman[1]*

1 Department of Psychology, Dominican University, River Forest, Illinois, United States of America, **2** School of Education, DePaul University, Chicago, Illinois, United States of America, **3** Department of Psychology, Concordia University Chicago, River Forest, Illinois, United States of America, **4** Department of Psychology, College of DuPage, Glen Ellyn, Illinois, United States of America

* rcalinjageman@dom.edu

OPEN ACCESS

Citation: Sanchez C, Sundermeier B, Gray K, Calin-Jageman RJ (2017) Direct replication of Gervais & Norenzayan (2012): No evidence that analytic thinking decreases religious belief. PLoS ONE 12(2): e0172636. doi:10.1371/journal.pone.0172636

Editor: Michiel van Elk, Universiteit van Amsterdam, NETHERLANDS

Received: September 23, 2016

Accepted: February 6, 2017

Published: February 24, 2017

Data Availability Statement: All of our materials, raw data, and analysis files can be found on our project page on the Open Science Framework (https://osf.io/qc6rh/).

Abstract

Gervais & Norenzayan (2012) reported in *Science* a series of 4 experiments in which manipulations intended to foster analytic thinking decreased religious belief. We conducted a precise, large, multi-site pre-registered replication of one of these experiments. We observed little to no effect of the experimental manipulation on religious belief ($d = 0.07$ in the wrong direction, 95% CI[-0.12, 0.25], $N = 941$). The original finding does not seem to provide reliable or valid evidence that analytic thinking causes a decrease in religious belief.

Introduction

Religion seems to be a cultural universal and yet there are marked individual differences in degrees of religious belief and practice. One factor that might explain individual variation in religious faith could be a proclivity for intuitive styles of cognition over more analytic/reflective modes of cognition. This possibility was recently explored in Gervais & Norenzayan (G&N) in a paper published in *Science* [1]. Specifically, G&N reported a weak negative correlation between the tendency to engage in analytic thinking and belief in God. Moreover, G&N reported four experiments in which manipulations meant to increase analytic thinking substantially reduced self-reported religious belief.

Since publication, concerns have emerged about these findings. Specifically, an analysis of psychology papers published in *Science* flagged the paper by G&N [2] for failing a test for excess significance [3]. However, there is spirited debate about whether or not tests of excess significance can be meaningfully interpreted at the level of individual papers [4].

To provide an unbiased estimate of the degree to which manipulating analytic thinking affects religious belief we conducted a precise, large, pre-registered replication of Study 2 of

COMMENT · 20 MARCH 2019

Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

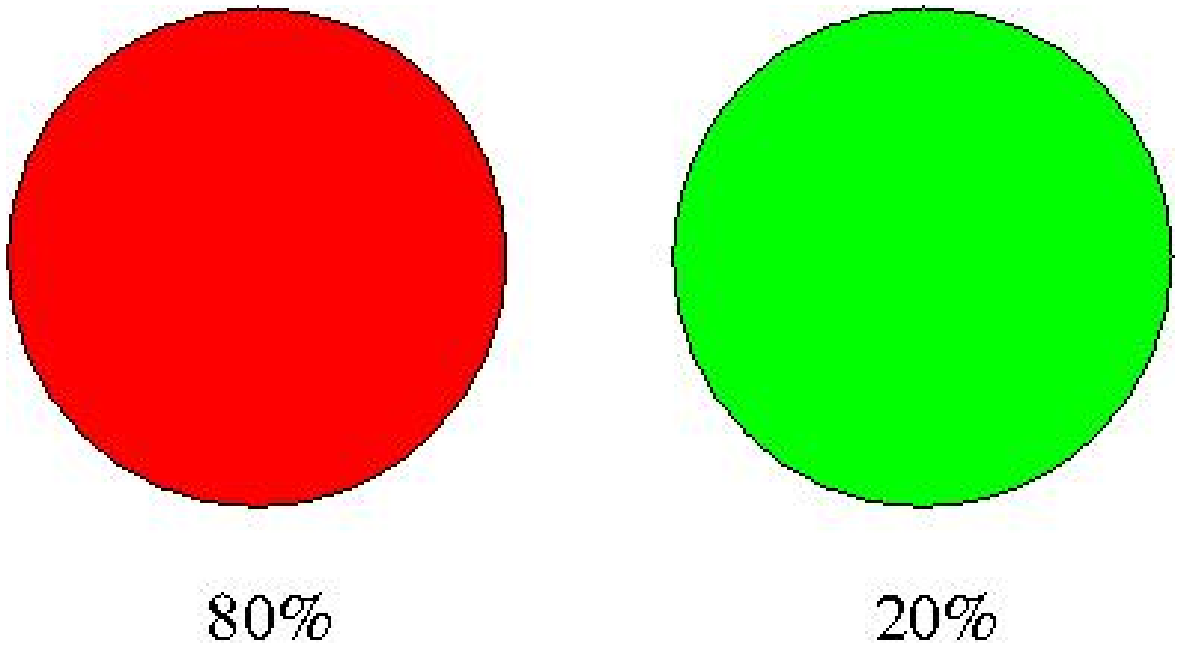
Valentin Amrhein ✉, Sander Greenland & Blake McShane

I nättidningen *Vox* den 22 mars 2019 förklarar och kommenterar Brian Resnick budskapet för en bredare publik. Hans avslutning:

One young scientist told us, "I feel torn between asking questions that I know will lead to statistical significance and asking questions that matter".

The biggest problem in science isn't statistical significance; it's the culture. She felt torn because young scientists need publications to get jobs. Under the status quo, in order to get publications, you need statistically significant results. Statistical significance alone didn't lead to the replication crisis. The institutions of science incentivized the behaviors that allowed it to fester.

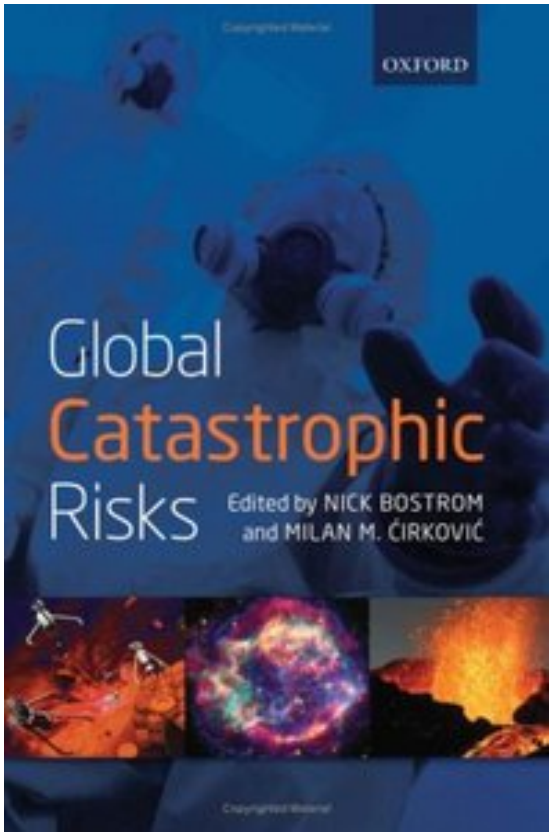
Ett mer välbelagt exempel: alltför "trigger-happy" mönsterdetektering.



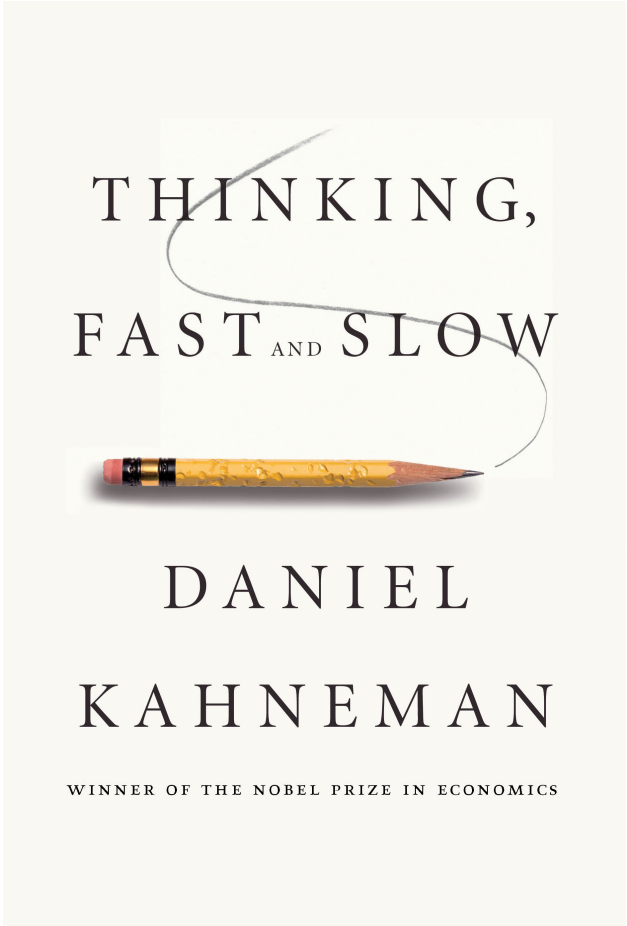
Råttor och duvor: Identifierar snabbt vilken lampa som lyser oftast, och gissar därefter rätt 80% av gångerna.

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.

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:



Se emellertid bloggposten **Reconstruction of a Train Wreck: How Priming Research Went off the Rails** av Ulrich Schimmack, Moritz Heene och Kamini Kesavan (2 februari 2017), och Daneil Kahnemanns respons, varur följande kan saxas:

"What the blog gets absolutely right is that I placed too much faith in underpowered studies. As pointed out in the blog, and earlier by Andrew Gelman, there is a special irony in my mistake because the first paper that Amos Tversky and I published was about the belief in the 'law of small numbers', which allows researchers to trust the results of underpowered studies with unreasonably small samples."

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:

Nästa exempel: Nu vill jag använda er som försökskanier.

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

Vad var, enligt SCB:s statistik, Uddevalla kommuns invånarantal 2018?

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?*


VOA

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?*

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?*

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?*

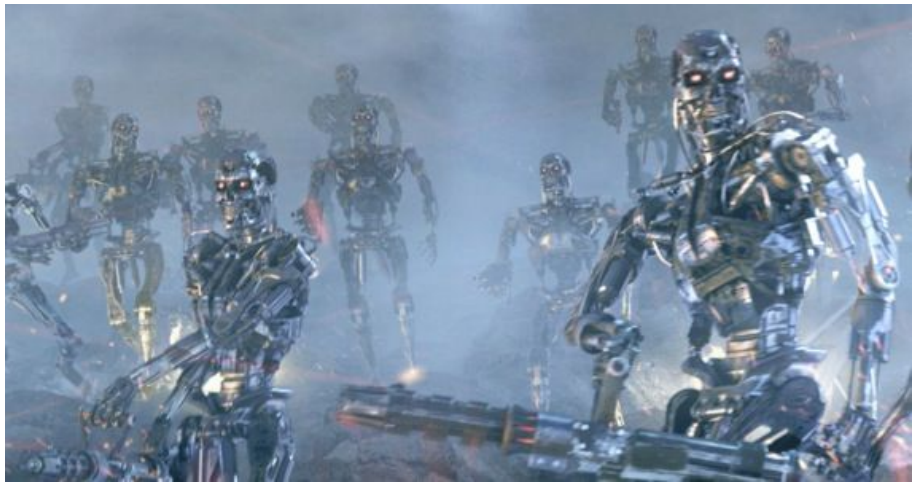
En annan grupp fick istället följande frågor:

> *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.

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*

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 (1) alltid betalar sig minst lika bra (och ibland bättre).

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".

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.

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.