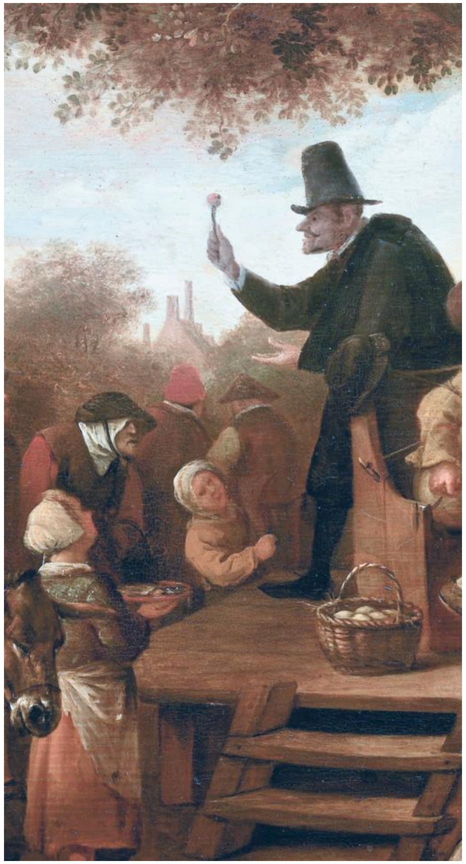
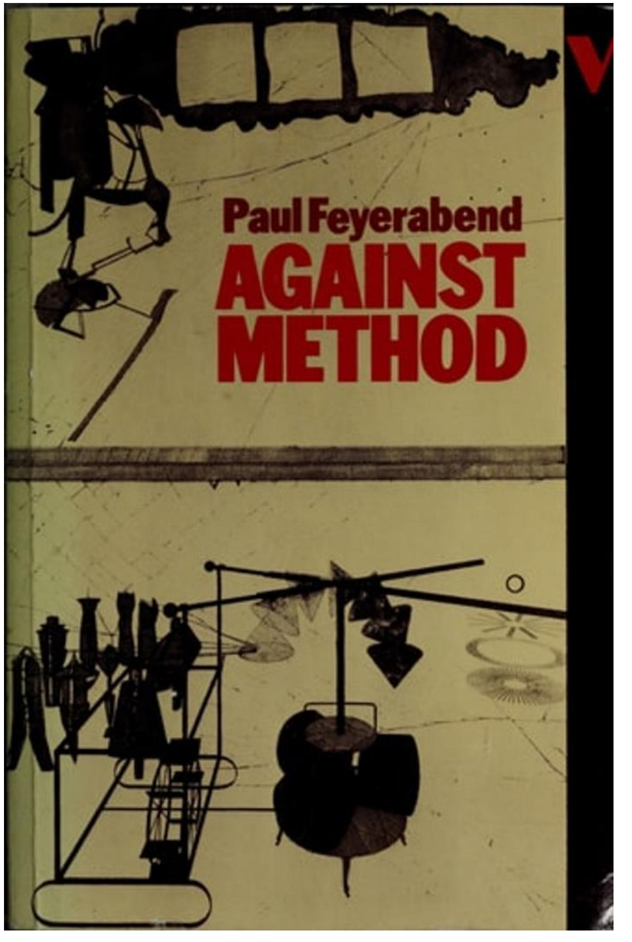


In defense of frequentism

Denny Borsboom
University of Amsterdam

Disclosure

- I regularly use Bayesian statistics as a practical Bayesian
- I have no objection to calculating and inspecting Bayes factors or posteriors
- I value various Bayesian ideas, e.g. on measures of evidence

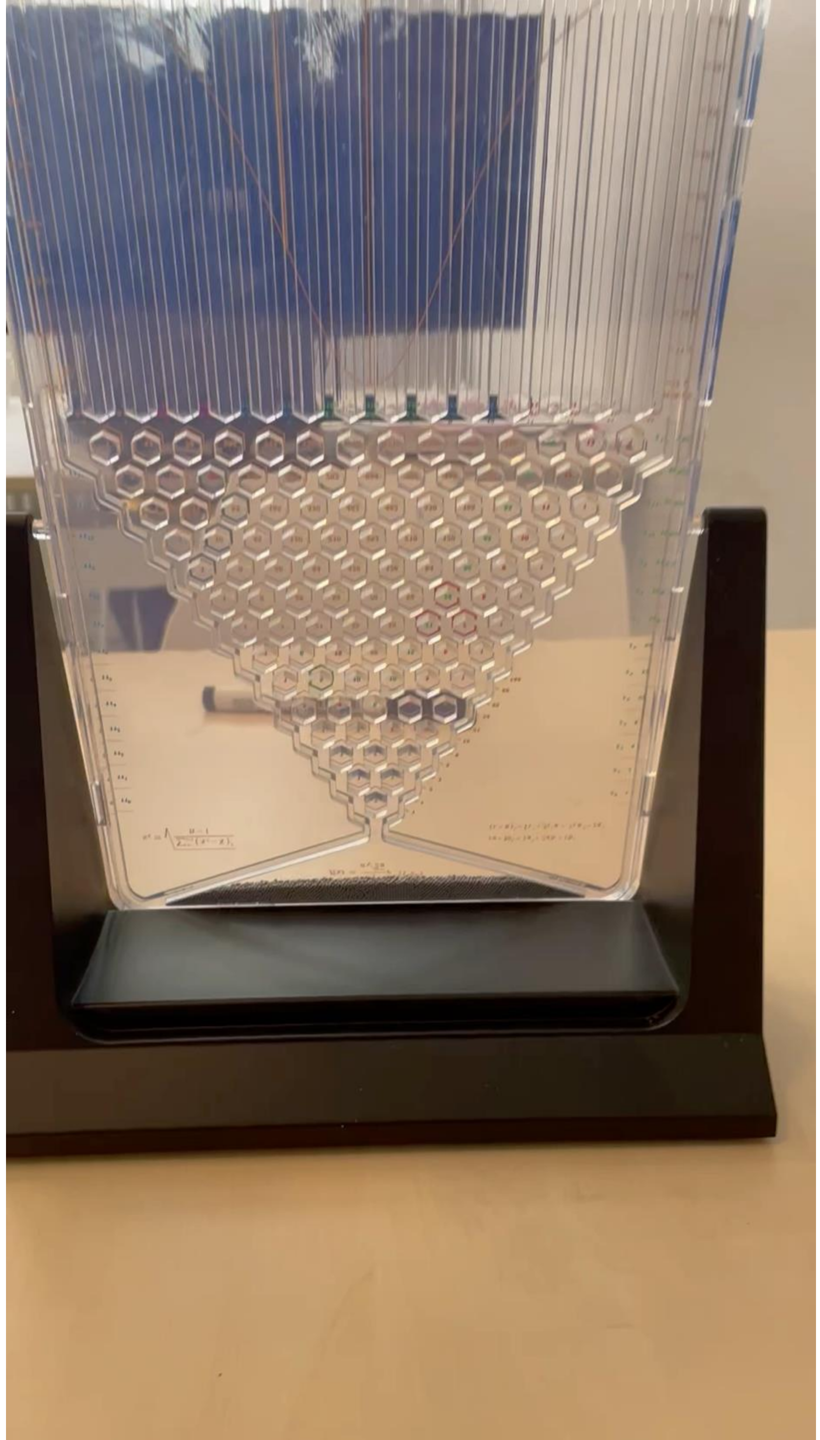


Borsboom, D., & Haig, B. D. (2013). How to practice Bayesian statistics outside the Bayesian church: What philosophy for Bayesian modeling? *British Journal of Mathematical and Statistical Psychology*, *66*, 39-44.

Overview

1. The frequentist conception of probability
2. Why I am not a Bayesian
3. The case for pluralism

Part I: The frequentist conception of probability



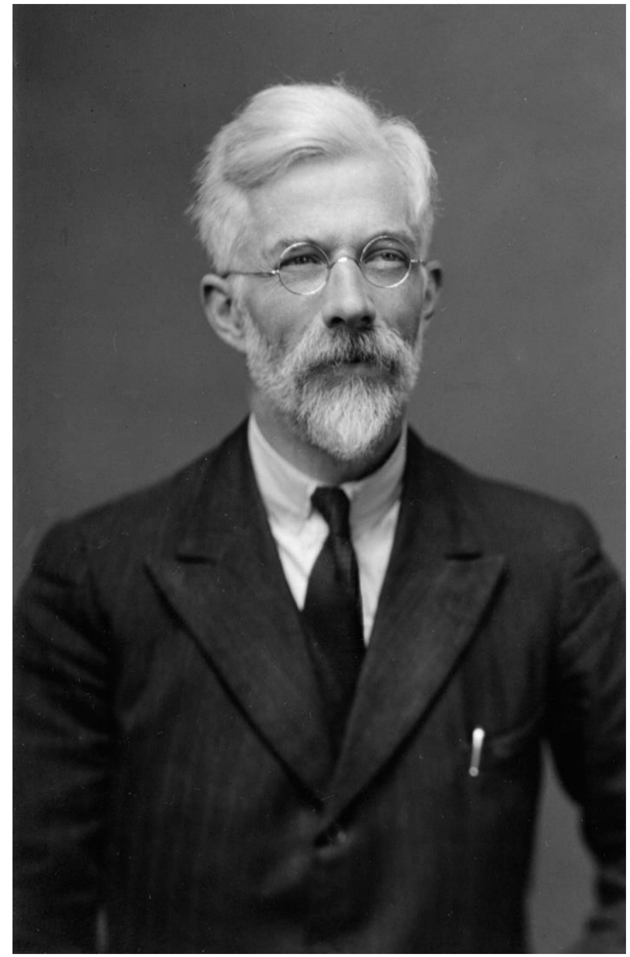
The mathematical characterization of probability

- Insight: Probability is the relative frequency with which an event occurs
- E.g.: “the probability of heads” means “the relative frequency with which a coin will fall heads”
- Letting the number of tosses approach infinity yields a mathematical limit
- This elegant theoretical construction is a conceptual masterpiece

Bartoš, F., Sarafoglou, A., Godmann, H. R., Sahrani, A., Klein Leunk, D., Gui, P. Y., ... Wagenmakers, E. J. (2025). Fair Coins Tend to Land on the Same Side They Started: Evidence from 350,757 Flips. *Journal of the American Statistical Association*, 1–10. https://doi.org/10.1080/01621459.2025.2516210

Probability and science: A spectacular combination

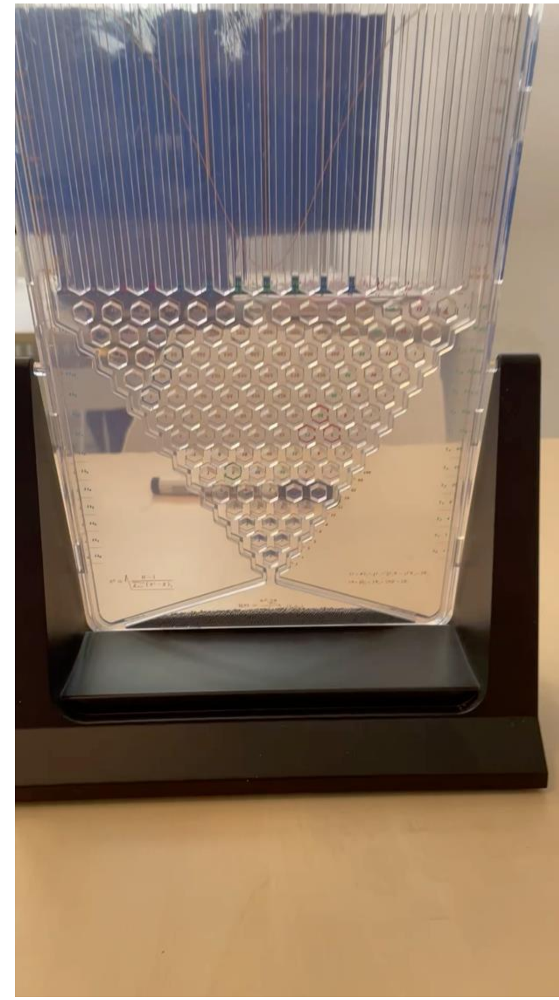
- Ronald Fisher had the insight that chance should not be *banned* from research design, but *used:*
  - Random assignment: letting chance allocate subjects to conditions
  - Random sampling: letting chance choose which elements from the population will be in your sample



Ronald Fisher (1890-1962)

Consequence

- *If* you use random sampling, *then* you *know* what the sampling distribution of your statistic is
- For instance, P(D|H) - the probability of a data D occurring given the truth of hypothesis H – equals the relative frequency with which D would be observed if H were true and we repeatedly drew samples of the same size as the original one

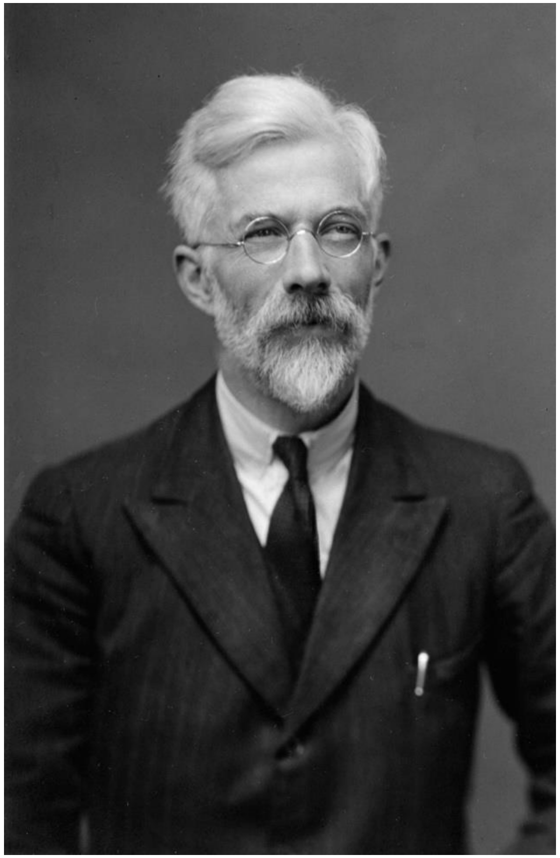


A powerful weapon

- P(D|H) - the probability of the data, given the hypothesis - is *real*
- It’s not an opinion, or the degree to which you should believe anything
- Given the explicit use of probability in research design (e.g. random sampling) there can be no argument about the values of frequentist probabilities
- In that sense (and *only* in that sense), frequentism is objective

Statistical inference

- By using P(D|H) judiciously, one can quantify uncertainty
- E.g., P(D>d|H) is the probability of observing data at least as extreme as d given H
- By using this fact, one can control the probability of Type I and Type II errors
- The standard null hypothesis test then guarantees at most 5% Type I errors



Ronald Fisher (1890-1962)

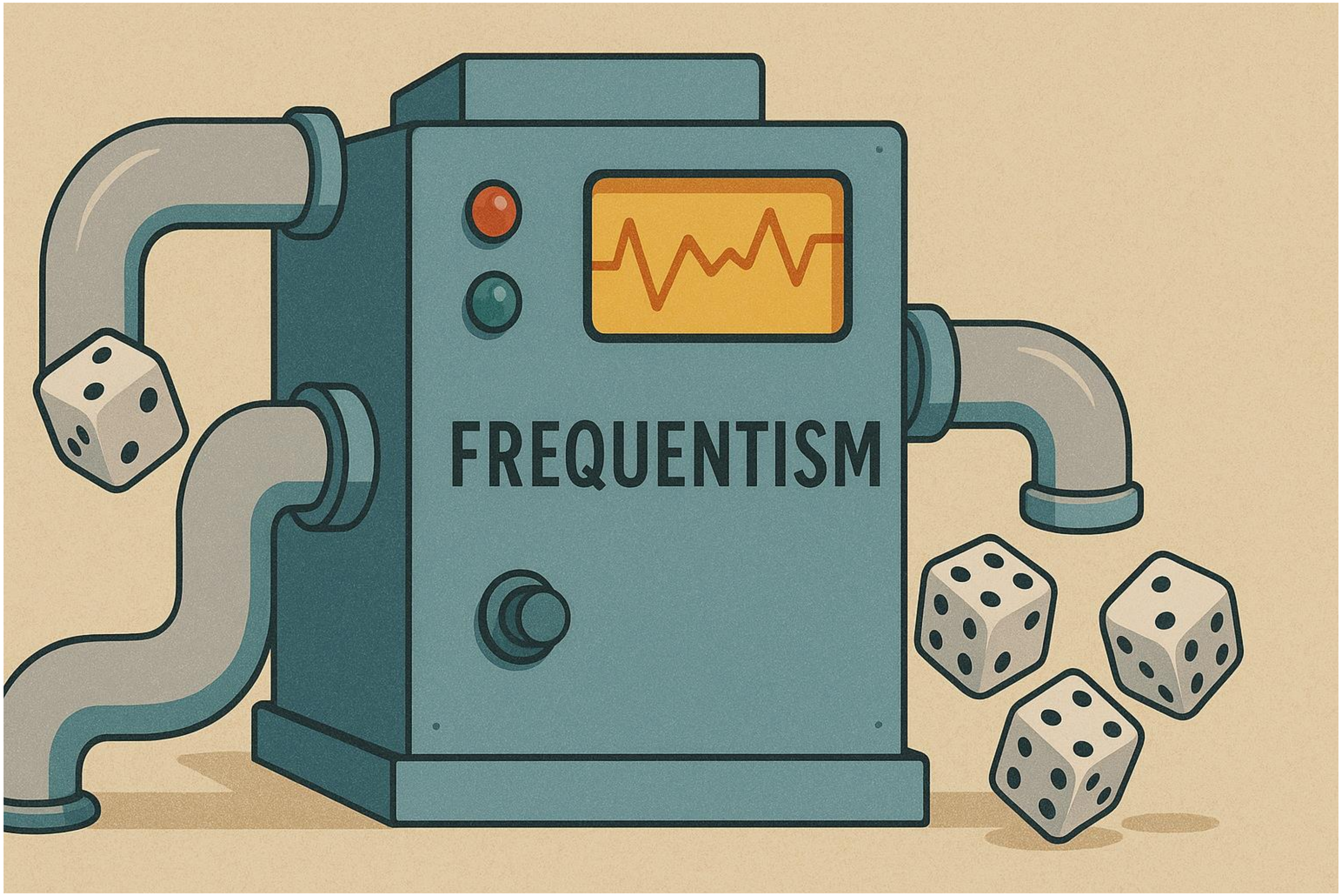
Advantages of the null hypothesis test

- Null hypothesis tests can be constructed for virtually all research designs
- The *p*-value *always* has the same interpretation
- Correct execution of tests *guarantees* 5% Type I errors at most
- The null hypothesis test rocks!

The argument structure of *p*

- Mary: “My intervention is effective, for I observed that the experimental and control conditions had different means!”
- John: “I think that’s just a coincidence.”
- Mary: “Well, if you were correct (so that H_0 is true), then the probability of observing this or a larger deviation from the null would be .01.”
- John: “Right. So if we were to repeat the experiment and the null hypothesis were true, then we would find such extreme deviations in 1% of the cases?”
- Mary: “Exactly right.”

Probabilities in, probabilities out



Thought experiments

- If random sampling or assignment actually happened, the probabilities are real
- If not, they are based on a thought experiment
- Such thought experiments build a semantic bridge for the application of probability
- Still useful, but limited

Borsboom, D., Mellenbergh, G.J. & van Heerden, J. Functional Thought Experiments. *Synthese* **130**, 379–387 (2002). https://doi.org/10.1023/A:1014840616403

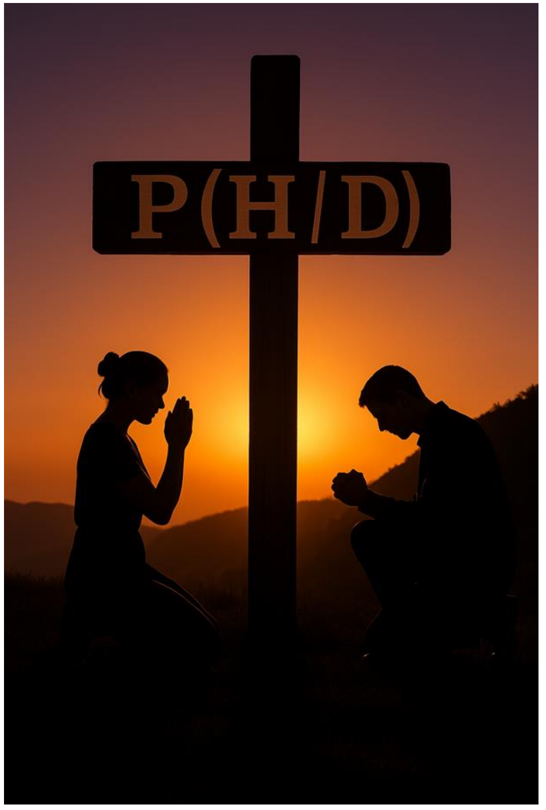
The social side of things

- As a social system, null hypothesis significance testing is certainly problematic
- This has to do with the sociology and psychology of science
- To believe that Bayes can fix this is naïve

Part II: Why I am not a Bayesian

My conversion attempt

- As noted, I am a part time Bayesian
- Some of my colleagues have been trying to convince me to expand this to full time Bayesianism
- Also Dennis Lindley apparently said that "Inside every Non-Bayesian, there is a Bayesian struggling to get out."
- Last Wednesday, I dedided to give my inner Bayesian a chance



The objective

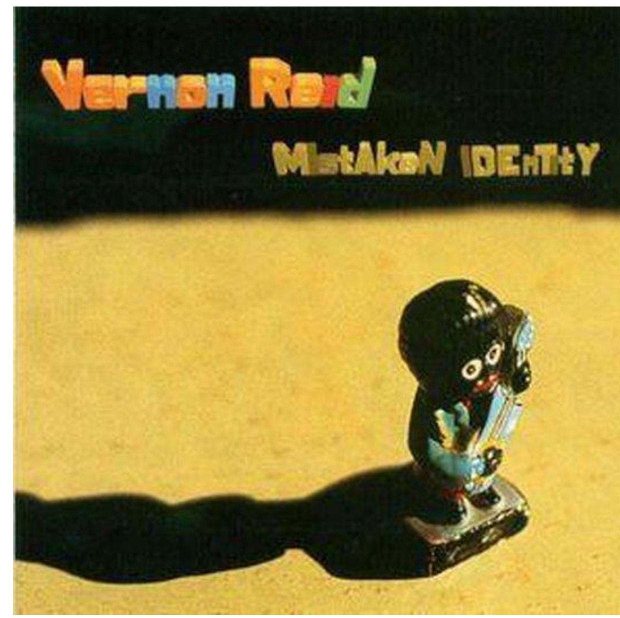
- Bayesians talk about how they can distill P(H) from the data:
- P(H|Data)≈P(Data|H)P(H)
- So, P(H|Data) now means "the degree of belief one does (or ought to?) attach to H, given the data"
- Or: P(H|D) tells me how to "rationally update my beliefs given the data"
- Or: P(H|D) allows me to "learn from data optimally"

My failure to be a true Bayesian

- My degrees of belief don’t seem continuous
- My degrees of belief are multidimensional
- My degrees of belief are influenced by pragmatic factors
- I don’t know how to update my degrees of belief
- I am not smart enough to understand how degrees of belief work in complicated models

My failure to be a true Bayesian

- My theories aren’t statistical models: P(Theory|Data) is not P(Statistical hypothesis|Data)
- I don’t limit my theory appraisal to predictive success and economy of the parameter space
- Instead, what impresses me about theories are things like:
  - Explanatory power
  - Plausible mechanisms
  - Strong analogies
  - Unifying force



My failure to be a true Bayesian

- And what about the quality of data?
- P(H|D) takes the data as given, but we all know there are good and bad researchers
- I assess evidence against ESP from Daryl Bem’s lab as inherently suspicious, but I take evidence from EJ’s lab quite serious
- How does data quality enter into the Bayesian scheme?

Journal of Personality and Social Psychology
2011, Vol. 100, No. 3, 407–425

0022-3514/11/$12.00 DOI: 10.1037/a0021524

Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem
Cornell University

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and premonition are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports 9 experiments, involving more than 1,000 participants, that test for retroactive influence by "time-reversing" well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur. Data are presented for 4 time-reversed effects: precognitive approach to erotic stimuli and precognitive avoidance of negative stimuli; retroactive priming; retroactive habituation; and retroactive facilitation of recall. The mean effect size (*d*) in psi performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results. The individual-difference variable of stimulus seeking, a component of extraversion, was significantly correlated with psi performance in 5 of the experiments, with participants who scored above the midpoint on a scale of stimulus seeking achieving a mean effect size of 0.43. Skepticism about psi, issues of replication, and theories of psi are also discussed.

Keywords: psi, parapsychology, ESP, precognition, retrocausation

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. The term is purely descriptive; it neither implies that such phenomena are paranormal nor connotes anything about their underlying mechanisms. Alleged psi phenomena include *telepathy*, the apparent transfer of information from one person to another without the mediation of any known channel of sensory communication; *clairvoyance* (sometimes called *remote viewing*), the apparent perception of objects or events that do not provide a stimulus to the known senses; *psychokinesis*, the apparent influence of thoughts or intentions on physical or biological processes; and *precognition* (conscious cognitive awareness) or *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process.

Precognition and premonition are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports nine experiments designed to test for such retroactive influence by "time-reversing" several well-established psychological effects, so that the individual's responses are obtained before the putatively causal stimulus events occur.

Psi is a controversial subject, and most academic psychologists do not believe that psi phenomena are likely to exist. A survey of 1,100 college professors in the United States found that psychologists were much more skeptical about the existence of psi than were their colleagues in the natural sciences, the other social sciences, or the humanities (Wagner & Monnet, 1979). In fact, 34% of the psychologists in the sample declared psi to be impossible, a view expressed by only 2% of all other respondents. Although our colleagues in other disciplines would probably agree with the oft-quoted dictum that "extraordinary claims require extraordinary evidence," we psychologists are more likely to be familiar with the methodological and statistical requirements for sustaining such claims and aware of previous claims that failed either to meet those requirements or to survive the test of successful replication. Several other reasons for our greater skepticism are discussed by Bem and Honorton (1994, pp. 4–5).

There are two major challenges for psi researchers, one empirical and one theoretical. The major empirical challenge, of course, is to provide well-controlled demonstrations of psi that can be replicated by independent investigators. That is the major goal in the research program reported in this article. Accordingly, the

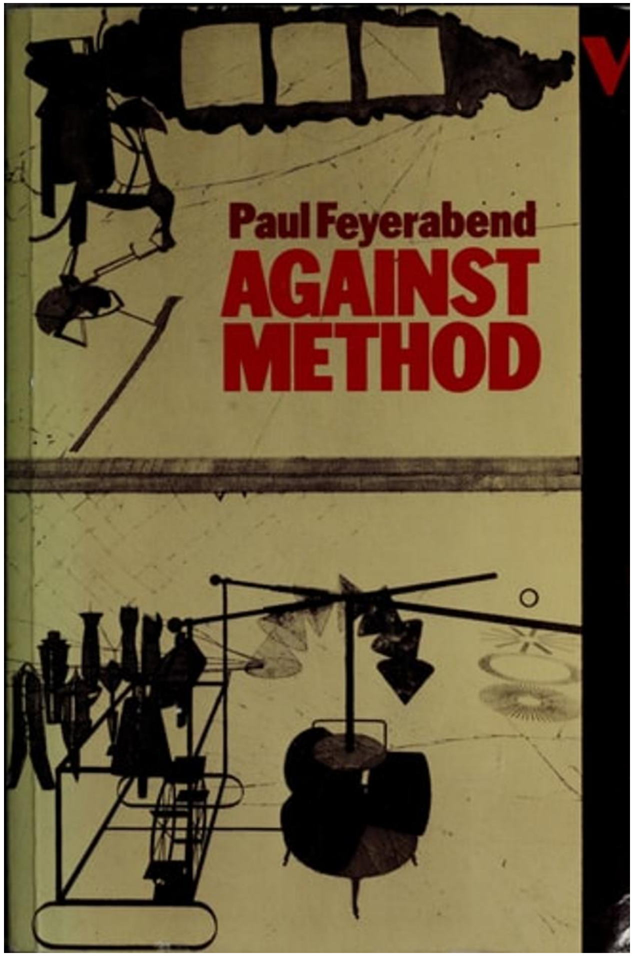
This article was published Online First January 31, 2011.

I am grateful to the students who served as head research assistants and laboratory coordinators for their enthusiasm and dedication to this controversial enterprise: Ben Edelman, Rebecca Epstein, Dan Fishman, Jamison Hahn, Eric Hoffman, Kelly Lin, Brianne Mintern, Brittany Terner, and Jade Wu. I am also indebted to the 30 other students who served as friendly and reliable experimenters over the course of this research program. Dean Radin, senior scientist at the Institute of Noetic Sciences (IONS), and David Sherman, professor of psychology at the University of California, Santa Barbara, provided valuable guidance in the preparation of this article.

Correspondence concerning this article should be addressed to Daryl J. Bem, Department of Psychology, Uris Hall, Cornell University, Ithaca, NY 14853. E-mail: d.bem@cornell.edu

Against statistical doctrines

- I think my primary problems with Bayesianism aren't technical
- Rather my objection is that Bayesianism wants to do *too much*
- Bayesian statistics should be seen as a method among many, not as a doctrine one converts to



Part III: The case for pluralism

What is the real problem?

The garden of forking paths

- Any statistical problem can be approached from a multitude of angles
- P-values, E-values, BF-values, etc.
- The choice between approaches should be seen as a robustness problem

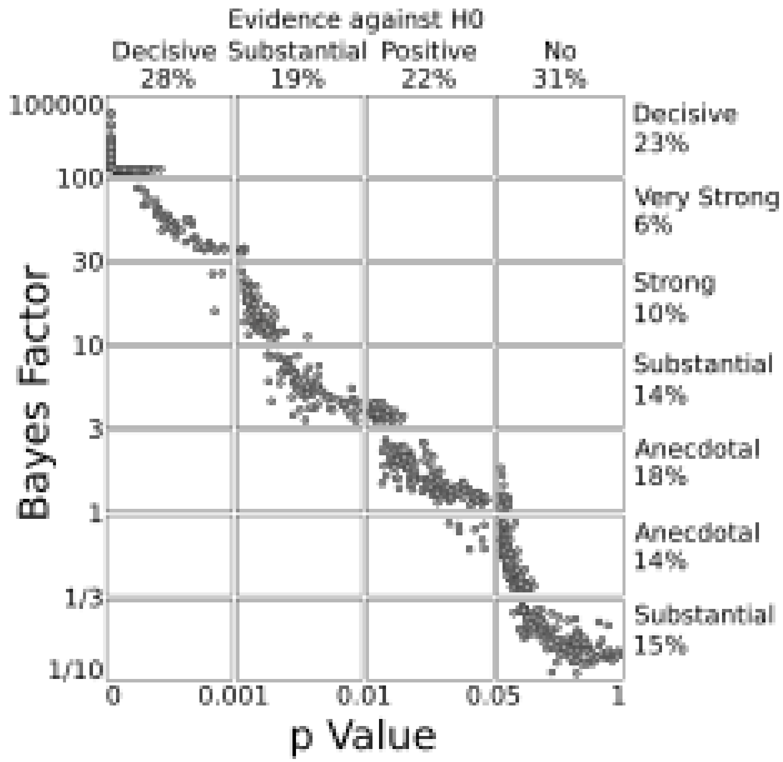
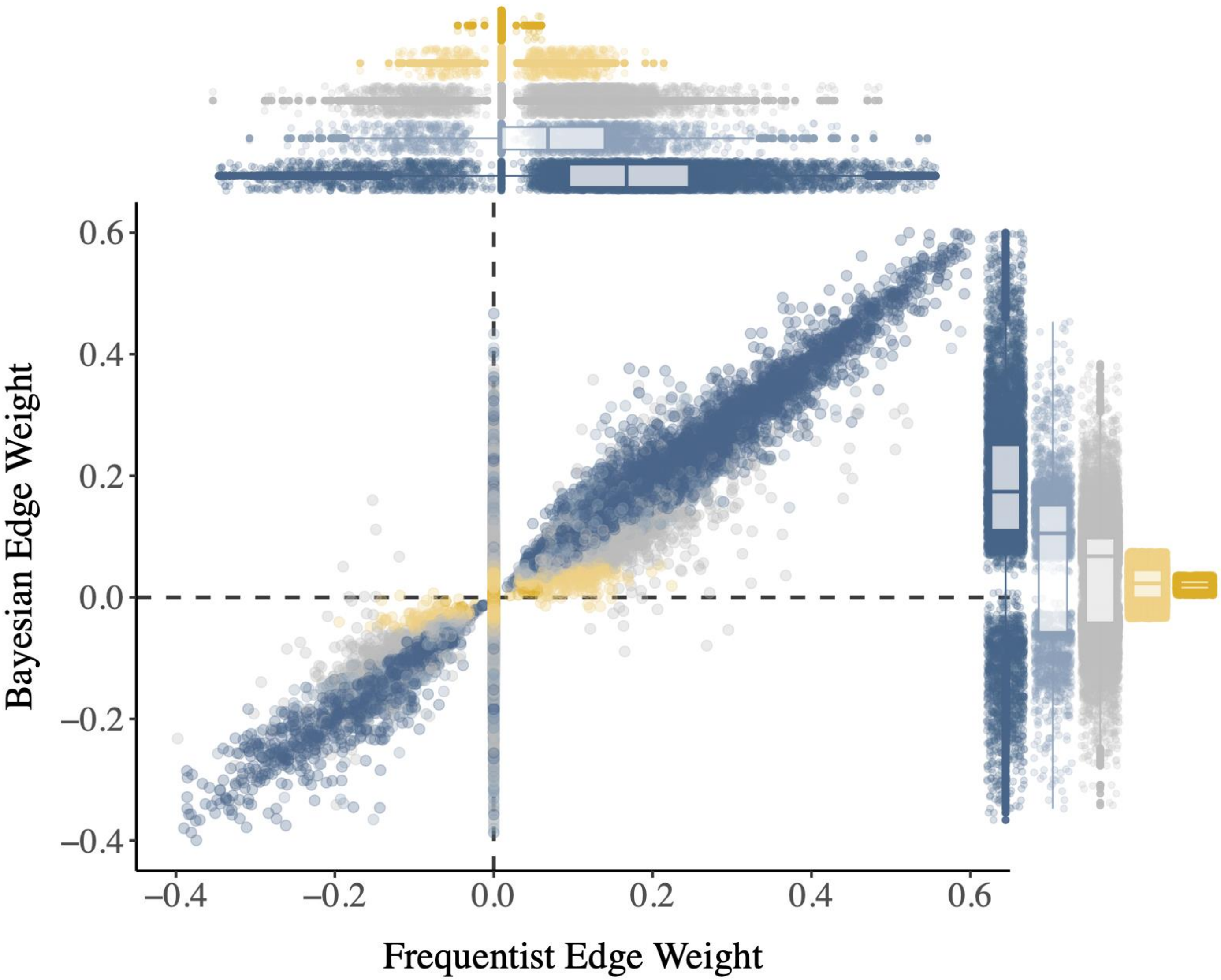


Figure 6.3: The relationship between Bayes factor and p value. Points denote comparisons (855 in total). The scale of the axes is based on the decision categories, as given in Table 6.1 and Table 6.3.

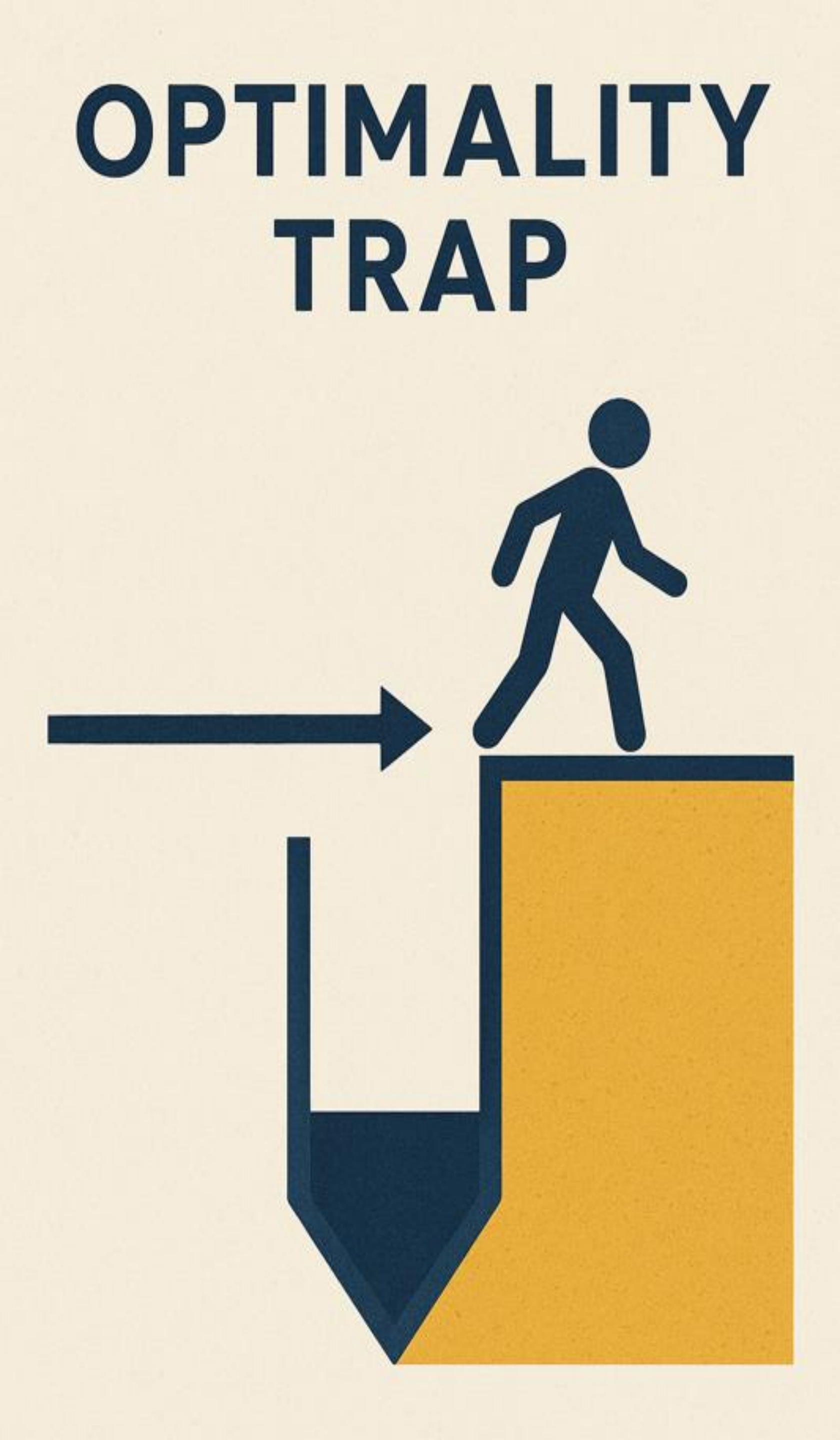
Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests. *Perspectives on psychological science : a journal of the Association for Psychological Science*, *6*(3), 291–298. https://doi.org/10.1177/1745691611406923



Huth, K.B.S., Haslbeck, J.M.B., Keetelaar, S. *et al.* Statistical evidence in psychological networks. *Nat Hum Behav* (2025). https://doi.org/10.1038/s41562-025-02314-2

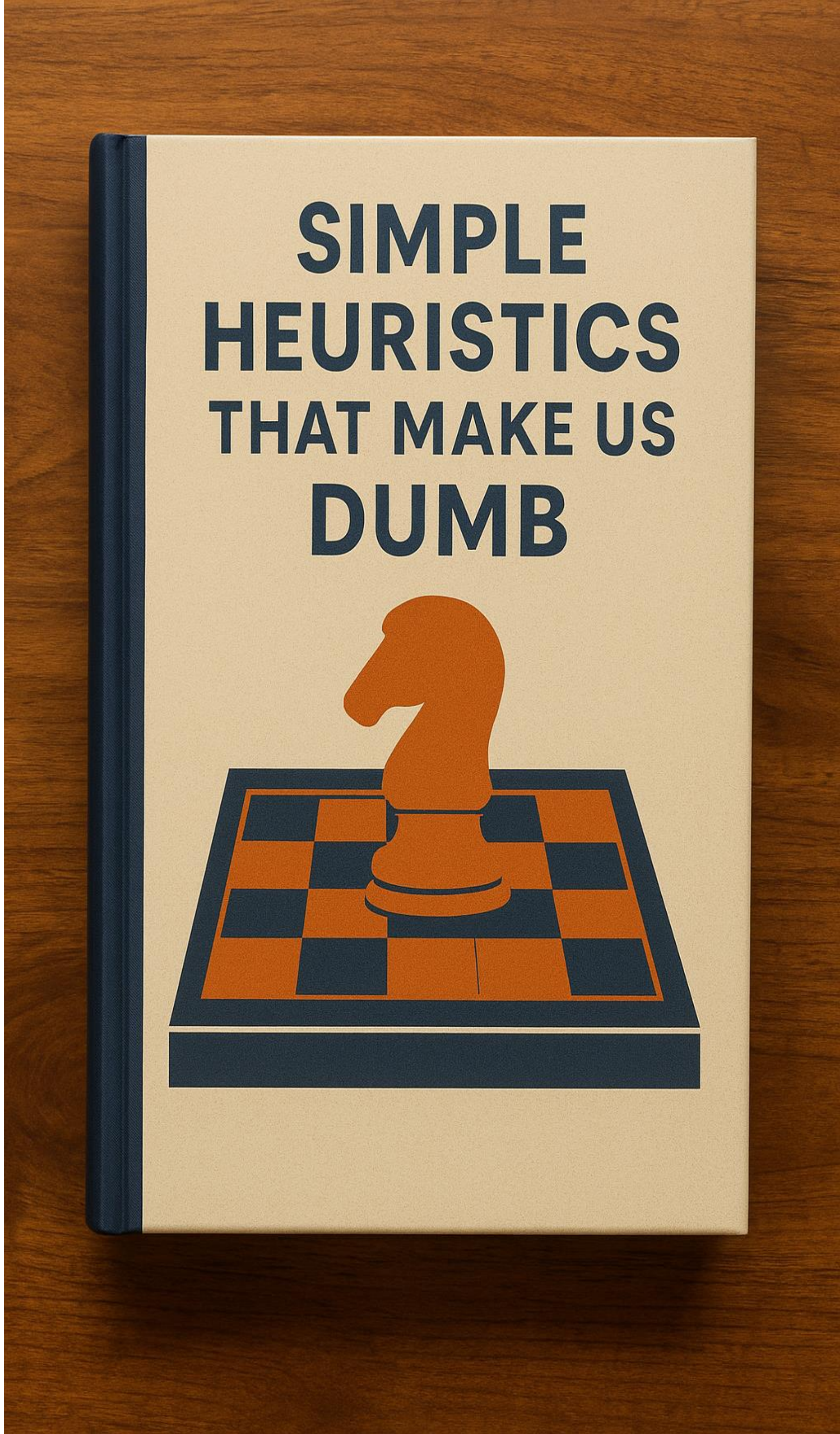
The optimality trap

- Many people look for "the optimal analysis"
- Statisticians are often asked to provide this
- We all know that there is no such thing
- In effect, we are limiting our view
- Missionary Bayesians and frequentists fall prey to tunnel vision



Simple heuristics that make us dumb

- A scientist drawing an inference on data is in a complicated situation
- No working scientist really understands the depths of statistics
- So: they are looking for a simple heuristic
- *BF>5 is exactly as problematic as $p<0,05$*

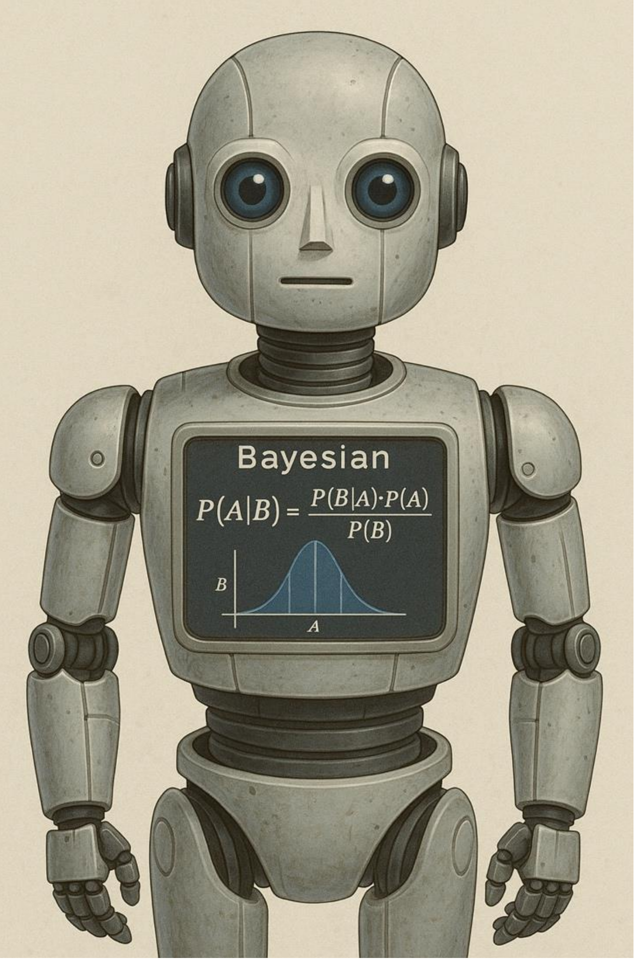


Magic bullets

- The *p*-value certainly has downsides: we all agree it is often misinterpreted
- But will Bayesian results not face the same problem? They are *at least as complicated*
- Sure, Bayesian techniques feature important ideas and have their place..
- …as tools among other tools
- But they will not be magic bullets
- The problem, after all, is not statistics but the human condition

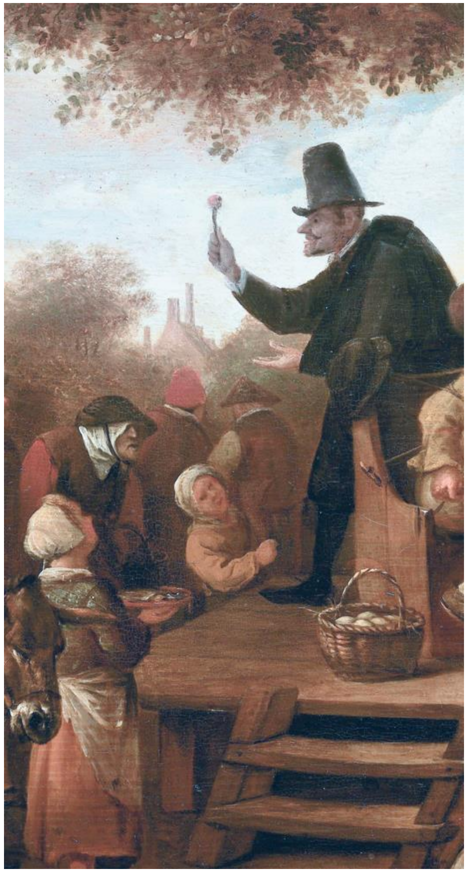
The pragmatic Bayesian

- In the words of Edwin Jaynes, Bayesian inference describes how a robot would do inference
- That is in many cases a useful perspective
- However, it's just one of many alternative perspectives
- So: look at Bayesian inference as one of many possible strategies



Conclusion

- Frequentist tools are limited, but useful in certain situations
- Bayesian tools are *also* limited, but useful in certain situations
- However, Bayesian statistics are often sold as remedies for poor *use* of statistical tools
- This is not just wrong, but silly
- We need to teach people to reason about different methods, not trade one mindless mouseclick for another



De kwakzalver, Jan Steen (ca. 1655)