ELSEVIER

Journal of Economic Behavior & Organization
Vol. 39 (1999) 29–40

JOURNAL OF
Economic Behavior
& Organization

A message from psychologists to economists: mere predictability doesn't matter like it should (without a good story appended to it) [☆]

Robyn M. Dawes [*]

Department of Decision Sciences, Carnegie-Mellon university, Pittsburgh PA 15213-3890, USA

Received 11 November 1997; received in revised form 15 June 1998; accepted 20 July 1998

Abstract

Stephen J. Gould (Pittsburgh, 3/3/97) recently defined humans as 'the primates who tell stories.' This paper reviews evidence for a more radical definition as 'the primates whose cognitive capacity shuts down in the absence of a story' when attempting to incorporate probabilistic information to make a coherent probabilistic inference. Thus, people cannot conform ('descriptively') to the standard expected utility (EU) model of economic decision making, given that probabilities often cannot be combined either implicitly or explicitly in the absence of a good, clearly relevant story justifying the combination. Moreover, that inability severely limits the standard EU model for use in prescriptive decision making. ©1999 Elsevier Science B.V. All rights reserved.

People have a great deal of difficulty appreciating statistical contingency in the absence of a causal story that makes the contingency 'reasonable.' In particular, people often do not understand on a purely statistical basis the relationship between properties that describe sets versus those that describe members of these sets, particularly when thinking about human behavior. For example, people often 'underutilize' base rates—that is, the simple extent of the sets or classes about which inference are to be made—in the absence of a causal story; thus, in deciding whether a particular blue taxi was responsible for a particular hit-and-run accident in visibility poor enough to preclude definite identification of the taxi color, people will often ignore the base rate information that 90 percent of the accidents are due to blue taxis because 90 percent of taxis in town are blue; when, however, people are told that 90 percent of accidents are due to blue taxis *because* the blue taxi company hires inferior drivers and does not train them well, this base rate information is incorporated into judgments about the particular accident (given it creates a story about why we should

[☆] This paper was first presented at a conference of Psychology and Economics, Faculty of Commerce and Business Administration, University of British Columbia, Vancouver, BC, Canada, 6/6/97–6/7/97.

[*] Fax: +1-412-268-6938.

PII: S0167-2681(99)00024-4

suspect a particular blue taxi; see, e.g. Ajzen, 1977). Conversely, even some professorial researchers in psychology fail to understand that base rate inference is based on a positive correlation between overall frequency and individual outcomes, and that because correlation is symmetric, it is possible to infer base rates from single observations; that inference has been termed a 'false consensus effect' if the observation happens to be about one's own behaviors or predilections or attitudes. In contrast, people will accept a causal story of a single incident or event as convincing, even if there is absolutely no general contingency established between the elements that are claimed to be 'causal' in the story and those that are claimed to follow from these causes.

I chose to oversimplify in the statements I have just made, because some people some of the time do accept some contingency without causality. For example, on hearing that in general one type of medical procedure for a particular problem is twice as successful as another type, we generally opt for the former type—even in the absence of any understanding whatsoever of the details of the procedures and the conditions. It works. I want it. In such medical contexts, we accept as quite natural the inference from the group of people (taking a vaccine, having the operation, ingesting the drug) to us as a single individual—or to our children. In contrast, however, most of us do not accept the statement that since most people in our income category vote Republican (or Democrat) we also are likely to vote Republican (or Democrat) in the future. Our political beliefs are clearly influenced by factors that have nothing to do with our incomes, and many of us would find it almost insulting to be told that whether we know it or not we are 'influenced' by our financial status. Most of us regard our political beliefs as 'basic attitudinal baggage,' (Abelson, 1986, 1988) to which we are committed for deep philosophical and personal reasons. Even if we were told that the inference from income to individual party identification was *stronger* than the inference concerning the success of the medical procedure, we would find this former inference (many of us would anyway) strange and unacceptable. Pure financial concerns cannot 'cause' our beliefs. On the other hand, there must be something about basic human physiology that led one treatment to be twice as successful as the other, and that would therefore 'cause' it to be more likely to be successful for ourselves.

The point, however, is that whether we are consistent or not, our failure to rely on a probabilistic contingency when it is not embedded in a causal story violates standard neoclassical economic theory. That theory is based on the valence (either inferred or directly estimated) of outcomes and their probabilities (objective or subjective). It does not matter how the probabilities come about. An estimated probability is an estimated probability, which will affect the estimate of the expected utility of an outcome. The same estimated probability should not affect our behavior dependent on whether or not it is wrapped up in a causal story. But it does.

For example, base rate neglect is well known. In fact, it is so well known that now a common way to attempt to make a mark in the field of behavioral decision making is to state that it is overblown (Koehler, 1996), to find some context in which it does not occur, or even to question the applicability of standard probability theory: for example, by claiming that probabilities cannot be applied to single events as opposed to collections of events (Gigerenzer, 1996), or that only relative frequencies make sense psychologically (Cosmides and Tooby, 1994), or even that it is by definition impossible to make an 'epistemic' error in assessing probabilities—because like Popeye we are what we are, although 'ontologically' the laws of probability may be valid even if no one existed (Cohen, 1981).

That base rate neglect is alive and well can be observed in the current debate in clinical psychology over 'recovered repressed memory,' often leading to the discovery of remarkable and severe abuse in satanic cult situations. The argument is that relatively high base rate distressing conditions—such as a negative physical self- image, or an eating disorder—can be diagnostic of an extremely low base rate 'causal mechanisms' (such as being raised in a satanic cult, which may have base rate probability zero). That is just flat out irrational.[1] Diagnosing this causal condition on the basis of the problem is equivalent to claiming it is a necessary condition for the problem to occur. But it cannot logically be a necessary condition when its base rate is so much lower. Nevertheless some clinicians write that it is even *unethical* not to 'search for possible etiology involving satanic cult ritual abuse or at least incestuous sexual abuse' when finding problems such as a negative self-image or eating disorders. (In fact, there is not even any evidence that the inference works in the other direction either, but that is another matter.) Note, however, how compelling the 'story' is. Suppose you have been abused sexually day after day after day when you are an infant or child and then hypnotized to forget it, and perhaps even programmed by the satanic cult people to turn to their beck and call if you receive certain 'triggering messages' (such as a postcard saying we love you and look forward to seeing you—actual example); then *of course* you would not feel very good about your body, and you may seek solace in excess eating (or drinking or ingestion of legal or illegal drugs).

The base rate experimental findings are quite consistent with the causal story idea. The fact that 90 percent of the taxis in an area are blue does not appear to affect the judgment about whether a particular taxi that might be green but could possibly be blue is in fact blue or green. On the other hand, the story that 90 percent of the accidents are caused by blue taxis because the blue taxi company does not screen its drivers well or train them adequately does affect the judgment. Causality (Ajzen, 1977; Bar-Hillel, 1980) makes the base rate relevant, and may even lead to absolutely bizarre judgments. For example, if people dream in color with probability 0.80, and if there is no relationship between the types of dreams of pairs of sexual partners, then in 64 percent of the pairs both will dream in color, in 4 percent neither will dream in color, and in 32 percent one will dream in color and one will not. That follows from elementary probability theory; $0.8 \times 0.8 = 0.64$; $0.2 \times 0.2 = 0.04$, and $0.2 \times 0.8 + 0.8 \times 0.2 = 0.32$. After being told, however, that 68 percent of pairs are concordant and that one dreams in color, judges believe that the probability that the other dreams in color as well is 0.68, not 0.80. Somehow, it is easy to construct a story that 'like seeks like'—and that the type of dreams people have reflects some type of deep psychological characteristics that others somehow appreciate when seeking partners. At least, it is easy to make up such a story in a society that has yet to wean itself from the fantasy life of Sigmund Freud.

Note that what is happening in the examples is that the probabilistic information easily available to the judges is not incorporated into a final probabilistic judgment. When such incorporation fails, then the final probability judgment is adversely affected. Each component prior to combining information is, however, not affected. In contrast, a story, which

[1] Since $P(A/B) = P(A \text{ and } B)/P(B)$ and $P(A \text{ and } B) \leq P(A)$, then $P(A/B) \leq P(A)/P(B)$.

provides causal information, creates a cohesion among the elements of the inference that can otherwise be lacking, with the result that some of these elements are underutilized or even ignored completely.

In fact, a cause or a good narrative often provides causal information that has an effect on all potential outcomes, rather than specifying a subset to which a probability applies. For example, we can imagine that in *general* 'like seeks like' in dreamers, and that selection and training may affect driving. Thus, as Bar-Hillel (1990) has demonstrated, people attend to base rate considerations that can possibly be construed (again, through a 'good story') to have some effect on all objects of judgment (in her terms, are incorporated into peoples' priors). In contrast, base rates that simply make it more or less likely to observe a particular instance, one whose characteristics are defined totally independent of the base rates, tend to be ignored. Yet, each type of base rate is equally important in reaching a rational judgment. Bayes's theorem does not distinguish between base rates that change probabilities by affecting the 'mix' of particular instances versus those that appear to 'nudge' all in one direction or another. A conditional probability is a conditional probability.

One way of explaining why there is a relationship between base rates and single occurrences is by pointing out that there is a correlation between them. This correlation can be computed by regarding one variable as a vector of zeros and ones and another variable as a vector of base rates. More ones will be associated with higher base rates, more zeros with lower base rates. In fact, it is very easy to figure out what the correlation should be. (See Dawes, 1989.) Moreover, it turns out that a standard regression analysis yields a result that the predicted value of the zero, one variable is *precisely* the base rate value with which it is paired.

But correlation is symmetric. Not so, according to some of our own colleagues—who argue that if there is a relationship between a subject's own response and her estimate of base rate responding, the subject has evidenced an 'egoistic' bias termed the 'false consensus effect' (Marks and Miller, 1987). The 'argument' made for the falsity and egoism of this effect is to point out that there is a positive relationship between the *error*—defined as the difference between the population value and the estimate—and the subject's own response. That is not a good argument, because there is *always* a positive correlation between an unbiased estimate of the population value and the discrepancy between that estimate and the true population value. How can that be when the estimate is 'unbiased?' The answer is that it is unbiased because its *expectation* is the population value, but the discrepancy between a *particular* value sampled and the actual population value *must in general* be in the direction of the value sampled. For example, consider that two of us attempt to assess the average weight of college students in a particular institution by randomly sampling five students. The actual average of our sample weights is an unbiased estimate of the mean population weight. But the discrepancy between this average and any constant value will be in the direction of the average, and in particular the discrepancy between the average and the actual population value will be in the direction of the average.

Let me express this conclusion in another manner. In the original study proclaiming the 'false consensus effect,' Ross et al. (1977) asked Stanford students to walk around the Stanford campus with a big sign board reading 'Repent!' Some agreed to do so and some declined. Subsequently, all of the subjects were asked to estimate the proportion of Stanford students who would agree. Those who agreed estimated on the average that about 62 percent would, while those who declined estimated that on the average 29 percent would. Well, clearly not everyone can be correct, and the direction of the error from the true proportion (even though we do not happen to know its value) is in the direction of the subject's own behavior. Hence, the 'false consensus.'

But consider this problem from a Bayesian perspective. Two people are asked to estimate the proportion of blue versus red chips in a book bag, and one person draws a blue chip while the other person draws a red chip. The best judgment of these people should not be equivalent. Consider, for example, that the judgments were the same; then the sampling of the single chip would not affect the estimate, which would mean that the prior belief following the first draw should be the same as that following no draws, which should then imply that the belief should not be changed by the second draw, and so on. Eventually a sample of 100,000 should have no effect on belief. (Happily, probability theory proclaims the same results whether we think of large samples in terms of the entire sample or in terms of successive samples of size 1; e.g. readers may wish to convince themselves that the probability of drawing 5 spades at random from a deck of cards is $[13/52] \times [12/51] \times [11/50] \times [10/49] \times [9/48]$, which is exactly equal to the number of ways 13 things can be selected 5 at a time divided by the number of ways 52 things can be selected 5 at a time.)

As the believer in this false consensus effect may point out, my reasoning has to do with random drawing—but the subject herself or himself is not 'random.' True, but why is the subject for himself or herself any more or less diagnostic of what Stanford students would do than the previous subject would be? And certainly we would expect estimates to change on the basis of being told what someone else did, or we are once again in the problem of being unaffected by samples of any size whatsoever.

I started talking about this problem in 1989. What I want to convey in this paper is that many colleagues did not believe me until I provided examples involving small sample sizes. At that point, these colleagues conceded that I was correct but *only* for situations in which the subject constituted a 'significant' portion of the small group—because then the subject's own response would have a causal influence on the base rate in the group. When, I was told, I started dealing with large groups, my examples would show that there should be no inference at all based on a single sample of size 1. Well, of course not. In the first place the correlation between base rates and single outcomes is unaffected by sample size; also, denying that a sample of size 1 has any effect except in a finite sampling context would lead to the denial that a sample of any size would have an effect. The point is, however, that as I first talked about my own response as diagnostic, I used small sample statistics to demonstrate that the subject was *better off* attending to their own response than not, and the sophisticated psychologists to whom I talked immediately interpreted my results in causal terms and denied that they were in fact true when no causal explanation could be devised.

For example, consider Table 1, taken from my 1989 article, which present hypothetical 1,0 (e.g. 'yes'/'no') responses of three people to two items (e.g. 'I feel that my ideas may turn into insects,' and 'I look out for my rights.') Colleagues accepted my conclusion that to minimize mean square error (MSE) people should assign an estimate of 5/9 to responses they themselves endorsed and 4/9 to those they did not (leading to an MSE of 2/81), rather than assign the base rate 1/2 to each (leading to an MSE of 1/36). But then these colleagues claimed that this result held because 'there were *only* three subjects and two items' involved. (See App. A for a derivation of the 4/9, 5/9 estimates using both a Bayesian approach and

Table 1
Hypothetical Endorsements And Mean Squared Errors Of 4/9, 5/9 Estimates

		Items	
		1	2
(a) Response pattern			
People	1	0	1
	2	1	0
	3	0	1
(b) Squared error of	4/9, 5/9 responses		
People	1	1/81	1/81
	2	4/81	4/81
	3	1/81	1/81

a correlational one. Dawes (1989) has proved that these two approaches will always yield identical results for samples of size one.)

In addition to the logical analysis, we now have empirical data that people who do use their response as diagnostic are more accurate than people who do not (Dawes and Mulford, 1996). For example, when Oregon students were asked to estimate how the majority of Berkeley students responded to selected California Personality Inventory (CPI) yes/no items, they tended to be quite accurate overall when their estimates were consistent with their own response, but had no accuracy whatsoever when they predicted that the majority response would be different from theirs. Moreover, there was a positive correlation across subjects between the tendency to think that the Berkeley students agreed with their own responses and their accuracy, even partialing out the degree to which the subject's own responses were *actually* consistent with those of the Berkeley students. That result is consistent with previous findings of Hoch (1987). He found that people *under*weighted their own response in predicting general responses to a consumer survey (when these people were the consumers themselves, but not when these people were MBA students, whose responses—it turned out—were negatively related to those of the consumers). Hoch, however, believed that the false consensus effect was defined well *within* people, but still not across people, due to the direction of error argument. As I have shown elsewhere (Dawes, 1990), it is not possible for it to be valid within the but not between, or vice versa. *That does not mean that there is no such thing as a truly 'false consensus'—which consists of overweighting one's own response*, particularly in contrast to another person's response (Krueger and Clement, 1994). The point is that the classic definition of the 'false consensus effect' is flawed, and that when I first started pointing out the flaw, people believed me only when I could wrap my statistical argument inside a causal story.

Or consider another paradox. People suffer from the 'winner's curse' (Samuelson and Bazerman, 1985); this 'curse' is visited upon the winner of an auction, because the person who places the highest value on what is auctioned off wins it; but the fact that everyone else values it *less* often provides information that it may not be worth the amount of the winning bid. Now suppose that I win a Vickery auction for a rare postage stamp on which 20 other people bid (i.e. an auction where the winner pays the amount of the second highest bid plus a small increment). I have information that 20 other people—all of whom have at least enough knowledge of stamps to enter this auction—value the stamp less than I do. Should not that affect my valuation of the stamp negatively? On the other hand, suppose that I start haggling with a single stamp dealer over the same stamp, and the dealer gives in right away so that I buy it for the same price that I would have at the auction. The idea that 'there must be something wrong with the stamp' that could have caused the dealer to give in is a compelling one, even though a dealer who knowingly sold a flawed stamp as unflawed would risk expulsion from the American Philatelic Society and hence loss of dealership. Nevertheless, I may well have second thoughts.

But note the irrationality in terms of probability theory. In the latter case I have evidence that *one* person does not value the stamp as highly as I do, and one who wishes to *sell* it anyway, while in the former case I have evidence that 20 people do not, all of whom wish to *buy* it.[2] How is it that the latter case (i.e. one person's valuation) has a greater effect on my own valuation? The answer I propose is that it is very easy to make up a causal story about the single stamp dealer's willingness to reduce the price, while a story that involves 20 diverse other people is a difficult one to construct. Again, probabilistic assessment is not based on a rational consideration of probabilities per se, but rather on how this consideration can or cannot be wrapped within the causal story.

Another example concerns people's lack of appreciation that there is a high probability that some coincidences will occur in their lifetime, even though the probability that the particular ones that occur will in fact occur is quite low. As Gilovich (1991, pp. 176–177) writes:

> People fail to appreciate how many chances they have to experience something coincidental. Perhaps the key to this shortcoming of human intuition is that, unlike coin flipping, the repeated sampling is not obvious because it is not the *same distribution* being repeatedly sampled. By meeting a person here, thinking of someone there, receiving a phone call somewhere else, we are sampling from *different* distributions, and it is this difference that masks the repetitive element of the sampling process. Furthermore, people may be reluctant to think of their own experience, with all its attendant emotions, as a sample from a population of all possible experiences.

The point here is that when we think of tossing a coin again and again and again we have a causal mechanism that naturally leads to some embedded patterns that would be improbable if they were considered in isolation. People understand that, or at least some people understand that. The causal mechanism is easily visualized. In contrast, there are so many multiple factors going on that lead to the thought, the phone call, the premonition, or the lucky (or usually unlucky in the case of premonitions) guess that it is impossible to conceptualize a causal mechanism. The situation is a bit like the winner's curse versus the reduced demands of a negotiator. It is difficult to conceptualize all the different reasons why all those other people believe the stamp is less valuable than I do, but it is easy to postulate that a single negotiator must be anxious to sell because the stamp has some flaw that I do not notice.

Or consider how we assess what 'caused' airplane crashes. Do we do a statistical analysis by comparing crashes with successful landings to try to assess a contingency in this

[2] An alternative explanation in terms of expert dealers versus naive bidders does not account for the fact that most often many bidders are experts themselves, including some dealers.

comparison between some factor that we think might be involved in crashes and whether or not a crash occurred? No, not at all. We do a ‘fault tree’ analysis, often combining on a post hoc basis several factors that created the story involving *single* crash. We may do so by looking at the cockpit recordings of that crash. Do we compare them to the recordings of successful landings? No, in fact, those recordings are erased without analysis. Let me illustrate this example by quoting from an earlier article of mine (Dawes, 1993, pp. 14–16).

> Let me give a concrete example in which we believe that we understand a rather striking event quite well, yet the basis of our understanding could not accurately predict similar events. The event is the airplane crash of Western aircraft flight 903 at Mexico City on October 31, 1979. The plane landed at night on a runway that was under construction and crashed into a truck that had been left on that runway. The airplane crash occurred at 11:41:29 P.M. , which was the time the airplane was due to land. The left runway was closed to traffic, given it was the one under construction. The airplane went directly onto the left runway and into a truck on it.
>
> The first part of the FAA crash transcript (Airline Pilots Association, 1983) that strikes our attention occurs at 11:26:06, 15 min. before the crash. It is a cockpit conversation consisting of ‘Morning, Dan’ with a muffled ‘Morning’ in response. Dan was the navigator, who had a total of 4 hr sleep in the last 24, while the pilot had had 5. Later in the transcript we hear Dan say, ‘I think I’m gonna sleep late all night’ and ‘I think I got about 3 hours sleep this afternoon (at 11:31:51). Thus we can hypothesize that fatigue is an antecedent to the crash.
>
> The radar beam from the airport was on the left runway, while the right runway—but not the left—was illuminated with approach landing lights. The instructions were to come in on the beam (left) then shift to the right for the actual landing. The landing was in low visibility so that the construction on the left runway was not apparent. For example, at 11:30:38, the pilot says, ‘Yeah, smoke over the city’ and ‘Look at that smog.’ Thus, we can easily hypothesize that bad weather was an antecedent to the crash, for on clear night the construction might have been visible.
>
> Two minutes before the crash, the radio went dead. ‘What happened to that [expletive deleted] radio?’ the pilot asks. ‘Huh’ comes from the copilot. ‘The whole [expletive deleted] thing just, ah, quit; I don’t have any. . . ’ ‘It just died’. Here, we have an antecedent involving omission, specifically no radio contact 2 min. before the plane landed on the wrong runway.
>
> Sixty-five seconds before the crash, the control tower person stated, ‘twenty-six-o-five, you are to the left of the track’. By pure bad luck, the plane had been slightly to the left of the *left* runway. The pilot responded, ‘Yeah, we know’. ‘Just a little bit,’ the copilot added. Here, the problem of vague communication as an antecedent appears quite salient. Had the control tower person been explicit about the incorrect location of the airplane (e.g. ‘not the left runway, the right’) the crash might have been averted; ‘to the left of the track’ communicates an error that is broad indeed.
>
> Finally, 43 s before the crash, the control tower person confuses the two runways. ‘Ok sir, ok? Approach lights on runway 23 left but that runway is closed to traffic.’ In fact, the radar beam was on the left runway and the approach lights were on the right runway, which was not closed to traffic. Here we can easily determine another antecedent: stress interfering with clear thinking. Thirteen seconds later, the pilot realized that the plane

> was heading to the wrong runway but was unable to climb in the remaining 30 s to avoid the impact.
>
> There is a single consequent, the crash, and our perusal of the transcript leads us to find five antecedents: Fatigue, poor weather, communication breakdown (the radio), vague communication, and stress. All five combined for tragedy. Most probably, none of these alone would have led to the crash. Thus we are tempted to say that they are necessary conditions but not sufficient ones, until it is realized that the crash could have occurred following other antecedents. In addition, had something else happened (e.g., a slightly different location for the truck), there might have been no crash; hence, the antecedents are not sufficient either. But having the privilege of scanning what happened prior to the crash—a privilege that allows us in all such attempts to 'explain' an event in terms of what happened previously—we are able to pick these post hoc, and I would like to suggest that the choice is not a bad one at all. We are left with the generally accepted conclusion (M. Brenner, National Safety Board, personal communication, February 6, 1989) that in the absence of deliberate sabotage, most crashes occur as a result of a confluence of improbable events within a brief time frame.
>
> The point I want to make is that these factors would not allow us to predict future crashes very well at all. Airplane crews are often fatigued; bad weather occurs frequently; miscommunication is not that unusual, nor are temporary breakdowns of radio communication or panic at the last minute. (Back when airplane passengers were allowed to listen to air-ground communication, I once had the privilege of landing at O'Hare Airport while the controller was screaming at the pilot, 'I said runway 5, damnit, runway 5, 5 not 6, oh shit!')
>
> If we were, however, to do a *prospective* study of how well these precursors, either singly or in combination, predict whether or not a crash will occur, our measure of predictability (e.g., r, R^2, or percentage of correct classifications resulting from discriminate function, or whatever) would indicate gross unpredictability. The problem is that there is a many–many relationship between antecedents and consequences in the course of human life. As we retrospect, in contrast, we can create many–one relationships.

Yet another clear example of the preeminence of the causal story over 'mere statistical contingency' in social science can be found in the area of clinical psychology. Here, despite years of evidence that statistical (actuarial) prediction of important human outcomes is superior to clinical prediction (Meehl, 1954, 1986; Dawes, 1994, Chap. 3, Dawes et al., 1989), much of the practice of clinical psychology is based on the implicit assumption that clinical judgment must be superior. The superiority of statistical prediction is crystal clear when clinical judgment is pitted against actuarial analysis in a situation where both are based on the same information—so that the problem is basically one of how to combine it. It also has been found that—unlike some areas of business and medicine—clinical judgment in psychology is inferior in situations where the important variables captured by the statistical model constitute a proper subset of the variables considered by the clinician. (See Grove and Meehl, 1996 for the most recent review of the evidence.) It is also true that the statistical models need not even be optimal (Dawes, 1979; Dawes and Corrigan, 1974). Nevertheless, clinical psychologists make a great deal of money by relying on their intuitions for combining information and for making predictions, and in courts they eschew statistical models, instead proudly proclaiming that 'in my experience. . . .' What happens here is that the ‘inside view’ is preferred to the outside one (see Kahneman and Lovello, 1993), despite massive evidence that the outside one is superior.[3]

But can we adopt an outside view? An excellent example of the difficulty of doing so can be found in the ‘story based’ model of jury decision making (Pennington and Hastie, 1988). Basically, “recognition memory responses demonstrated that subjects spontaneously evaluated evidence in a legal judgment task by constructing an explanatory representation in the form of a narrative story. Furthermore, an item’s membership in the story associated with the chosen or rejected verdict predicted subjects’ ratings of its importance as evidence” (pp. 521). Further, stories are good or bad depending on how coherently they are narrated (only a few modern narrators’ employing the device of deliberately presenting information out of sequence to challenge the reader); hence, “The order manipulation shifted verdict choices in the direction of the more easily constructed story, implying that story structure causes decisions” (pp. 521). Believing in good stories is apparently what juries do (although ethical constraints require the use of mock juries, as opposed to manipulation of real ones). But what are they *supposed* to do? The job of a juror is to determine beyond a reasonable doubt that the defendant engaged in an illegal activity, or to determine that there is some reasonable doubt that the defendant did so. What possible relevance does the order of information and the coherence of the story have to such a determination? The answer is, of course, ‘none.’ Moreover, insight that a verdict might be manipulated by such order and story-telling abilities of defense or prosecution immediately yields a reaction among most potential jurors that such factors ‘should not’ affect the verdict. Knowledge of the possibility may lead to defense against it. It is unlikely, however, that counter measures can be taken on a purely intuitive basis ‘from the inside’ of the decision maker. Here, especially, some types of external aides (in the form of flow diagrams or pictures if nothing else) may be extremely helpful—*and* accepted once the power of the story in this context is clearly understood.

In closing, I want to point out that not everybody eschews ‘pure contingency’ in favor of stories in all contexts, even though they may often *believe* that they do. When I first discussed the airplane crash example with some engineering friends, they argued that their own retrospective analyses of this sort were perfectly appropriate, and moreover that they themselves had benefited greatly from them. The context within which they had benefited had been in the development of artificial heart valves, which they assured me are now quite safe. What had happened early in the development of these valves was that some people had developed blood clots and died. That problem was apparently related to the shape of the valve, as a retrospective, post hoc (fault tree) analysis suggested. So, my friends claimed that such analyses were wonderful. What they did not note but I did, however, was that

[3] For example, I quit clinical psychology after an expert interpreted a *single* response from a Rorschach Ink Blot test I had administered as meaning that the client was psychotic, or rather ‘pseudo-neurotic schizophrenic’; all the other Rorschach responses were ‘good form’ ones, which meant that the client actually had a higher proportion of such good form responses than is statistically usual for a normal population. But Card 8 does not look like a bear. My ‘supervisor’ held up that card at a staff meeting, explained to those present that I was an expert in measurement but did not understand people very well, and then challenged: ‘Does this look like a bear to you?’ As I mentioned, Card 8 of the Rorschach Ink Blot test does not look like a bear, and people agreed it did not. Well, what explained why the client saw a bear? The obvious answer was that she must have been hallucinating. Send her to the state hospital.

immediately after doing the post hoc fault tree analyses about how the shape might have caused blood clots, they conducted a series of experiments to see if a new shape did what they thought it would do. Given their training, they automatically accepted the idea of working prospectively to check out whatever good story they created. Such checking out was so natural to them that they did not even think about it, or at least think it was worth emphasizing. They were all impressed with their own causal stories, while I was impressed with the fact that they never stopped with them. I suggest, however, that in our own behavior and our own decision making in most economic contexts, we often 'stop here.' The result is that we do not appreciate probability without the story—which means that we could not possibly be expected to be optimizers, either on an implicit or an explicit basis, even after being trained to optimize.

Appendix A. A Bayesian Analysis of Table 1

Two items with endorsement probabilities 1/3, 2/3 are sampled with equal probability.

Sample a yes response. Then the posterior odds of sampling from the 2/3 endorsement item are 2 to 1, or the probability is 2/3.

Hence, the expected endorsement value is: $2/3 \times 2/3 + 1/3 \times 1/3 = 5/9$.

Sample a no response. Then the posterior odds of sampling from the 2/3 endorsement item are 1 to 2, or the probability is 1/3.

Hence, the expected endorsement value is: $1/3 \times 2/3 + 2/3 \times 1/3 = 4/9$.

Appendix B. A Correlational Analysis of Table 1

	Instances		Base rates	
Item no 1	1		1/3	
	0		1/3	
	0		1/3	
				$r = (1/36)/(1/2 \times 1/6) = 1/3$
Item no 2	1		2/3	
	1		2/3	
	0		2/3	
x	1/2	1/2		
s_d	1/2	1/6		

$\mathrm{Cov} = 5/18 - (1/2 \times 1/2) = 1/36$

Predicting base rates from instances

$p^1 = 1/3\ (I_i - 1/2) + 1/2$
$= 1/9\ I_i + 4/9$
$(= 5/9,\ 4/9 \text{ for } I_i = 1,\ 0)$

Predicting instances from base rates

$I^1 = 1/3\ (1/2)/(1/6)\ (p_i - 1/2) + 1/2$
$= p_i - 1/2 + 1/2$
$= p_i$

References

Abelson, R.P., 1986. Beliefs are like possessions. Journal for the Theory of Social Behavior 16, 223–250.

Abelson, R.P., 1988. Conviction, American Psychologist 267–275.

Airline Pilots Association, 1983 (May 16). Aircraft accident report: Western Airlines, Inc. McDonnell Douglas DC-10-10, N-903 WA Licenciado Benito Juares International Airport, Mexico City, D.F., October 31, 1979. Washington D.C. Authority.

Ajzen, I., 1977. Intuitive theories of events and the effects of base-rate information on prediction. Journal of Personality and Social Psychology 35, 303–314.

Bar-Hillel, M., 1980. The base-rate fallacy in probability judgments. Acta Psychologica 44, 211–233.

Bar-Hillel, M., 1990. Back to base rates. In: Hogarth, R.M. (Ed.), Insights in Decision Making: A Tribute to Hillel J. Einhorn, University of Chicago Press, Chicago, IL, pp. 200–216.

Cohen, L.J., 1981. Can human irrationality be experimentally demonstrated?. The Behavioral and Brain Sciences 4, 317–370.

Cosmides, L., Tooby, J., 1994. Better than rational: Evolutionary psychology and the invisible hand. AEA Papers and Proceedings 84, 327–332.

Dawes, R.M., 1989. Statistical criteria for establishing a truly false consensus effect. Journal of Experimental Social Psychology 25, 1–17.

Dawes, R.M., 1990. The potential non-falsity of the false consensus effect. In: R.M. Hogarth (Ed.), Insights in Decision Making. A Tribute to Hillel J. Einhorn University of Chicago Press, 179–199.

Dawes, R.M., 1993. The prediction of the future versus an understanding of the past: A basic asymmetry. American Journal of Psychology 106 (1), 1–24.

Dawes, R.M., 1994. House of Cards: Psychology and Psychotherapy Built on Myth, The Free Press, New York.

Dawes, R.M., Corrigan, B., 1974. Linear models in decision making. Psychological Bulletin 81, 95–106.

Dawes, R.M., Mulford, M., 1996. The false consensus effect and overconfidence: Flaws in judgment, or flaws in how we study judgment?. Organizational Behavior and Human Decision Processes 65 (3), 201–211.

Dawes, R.M., Faust, D., Meehl, P.E., 1989. Clinical versus actuarial judgment. Science 243, 1668–1674.

Gigerenzer, G., 1996. Our narrow norms and vague heuristics: A reply to Kahneman and Tversky. Psychological Review 103, 592–596.

Gilovich, T., 1991. How we know what isn't so: The fallibility of human reason in everyday life, The Free Press, New York.

Grove, W.M., Meehl, P.E., 1996. Comparative efficiency of informal (subjective, impressionistic) and informal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. Psychology, Public Policy and Law 2, 293–323.

Hoch, S.J., 1987. Perceived consensus and predictive accuracy: The pros and cons of projection. Journal of Personality and Social Psychology 53 (2), 221–234.

Kahneman, D., Lovello, D., 1993. Timid choices and bold forecasts: A cognitive perspective on risk taking. Management Science 39, 17–31.

Koehler, J.J., 1996. The base rate fallacy reconsidered: Descriptive, normative, normative and methodological challenges. Behavioral and Brain Sciences 19, 1–19.

Krueger, J., Clement, R.W., 1994. The truly false consensus effect: An ineradicable and egocentric bias in social perception. Journal of Personality and Social Psychology 67, 596–610.

Marks, G., Miller, N., 1987. Ten years of research on the false-consensus effect: An empirical and theoretical review. Psychological Review 102, 72–90.

Meehl, P.E., 1954. Clinical Versus Statistical Predictions: A Theoretical Analysis and Review of the Evidence, University of Minnesota Press, MN.

Meehl, P.E., 1986. Causes and effects of my disturbing little book. Journal of Personality Assessment 50, 370–375.

Pennington, N., Hastie, R., 1988. Explanation-based decision making: Effects of memory structure on judgment. Journal of Experimental Psychology: Learning, Memory, and Cognition 14 (3), 521–533.

Ross, L., Greene, D., House, P., 1977. The 'false consensus effect': An egocentric bias in social perception and attribution processes. Journal of Experimental Social Psychology 13, 279–301.

Samuelson, W.F., Bazerman, M.H., 1985. The winner's curse in bilateral negotiations. Research and Experimental Economics 3, 105–137.