Adaptive Probability Theory: Human Biases as an Adaptation

André C. R. Martins

Escola de Artes Ciências e Humanidades

Universidade de São Paulo

Av. Arlindo Bétio, 1000

Parque Ecológico do Tietê - Ermelino Matarazzo

CEP: 03828-080

amartins@usp.br

Phone: (55)(11) 6545.0243

Abstract

Humans make mistakes in our decision-making and probability judgments. While the heuristics used for decision-making have been explained as adaptations that are both efficient and fast, the reasons why people deal with probabilities using the reported biases have not been clear. We will see that some of these biases can be understood as heuristics developed to explain a complex world when little information is available. That is, they approximate Bayesian inferences for situations more complex than the ones in laboratory experiments and in this sense might have appeared as an adaptation to those situations. When ideas as uncertainty and limited sample sizes are included in the problem, the correct probabilities are changed to values close to the observed behavior. These ideas will be used to explain the observed weight functions, the violations of coalescing and stochastic dominance reported in the literature.

Keywords: *Rationality, Heuristics, Evolution, Bounded Rationality, Bayesian Inference, Adaptation, Weight functions, Decision Making*

1. Introduction

Suppose you are one of our ancestors, without any access to all knowledge mankind would accumulate in the next hundred thousand (or few million) years. It doesn't matter here if you are already a human or an ape or even some earlier ancestor. What matters is that you probably can't depend, when using your problem solving and decision making skills, on much more than what you observe and the little information you can obtain from the other members of your specie, by whatever communication means are available to you. If spoken language is not fully developed yet, the information you will get from your equals will most likely be very limited by human standards. Even if you are already a modern human, you would certainly not have access to all the information we have today and you have not learned anything about the rules of probability. But you must find ways to deal with uncertainty and, if you have ways to make predictions about the outcome of your actions, this might improve your survival chances. That is, if you have some way to analyze the world and come up with a description as close to the correct workings of your environment as possible, using your observations and what other people tell you about the frequency some events happen, that might prove to be a useful tool and a good adaptation, as long as the benefits you will obtain from it are greater than the costs associated with this tool, even if this tool is not perfect.

In other words, what you really need is a set of abilities that will allow you to obtain answers as close as possible to the correct evaluation, given your data. That is, answers close, but not necessarily equal, to those a full rational being would get. Of course, those abilities have to work well in the environment where you have to survive. If those abilities will fail in the laboratory tests your descendants will perform, but work

reasonably well in the complex problems you have to face, you are well adapted. Your descendants, when studying the mind they inherited from you might start to worry about what means to be rational and how and when their minds stop following the rules for rationality, but that is their problem. If getting closer to the correct rational behavior means more extra effort than the improved return from better decisions will provide, it won't be a good adaptation to do better than you are already doing.

But if you are one of the descendants, you might want to improve your decision making abilities, knowing what you should do and how you tend to think, including possible errors your brain is prone to make. And you will want to find explanations to why you actually think the way you do. While studying the problem of how a person should make a decision, Von Neumann and Morgenstern (1947) have shown it is possible to define rational behavior in a decision-making process as the behavior that chooses the alternative that will provide the highest expected utility return, given the probability distribution for the future events. Savage (1954) extended that result by introducing and developing the idea that what a rational being actually does is not maximize the expected utility by using a non-existent objective probability distribution, but maximize by using her subjective probability for the future events. Theses ideas have been developed into a full decision-making normative theory of rational behavior, where a rational agent will use Bayesian rules to update her knowledge and, given that knowledge, choose the best alternative (for a good review of many important results, see, per example, Bernardo (1994)).

This means, among other things, that rational agents must be able to analyze every piece of information available in order to obtain probabilistic evaluations of the possible consequences of their decisions, so that they can choose the correct action that maximizes their utility. In practice, a rational being would have to have a perfect

memory and be able to perform any required calculations in basically zero time, what is certainly impossible. Real humans, on the other hand, have to make their decisions using brains that are limited both in the speed and in the amount of information they can deal with. Departures from the normative utility-maximizing behavior should therefore be expected and there are several known examples of different “mistakes” done by people in their decision-making process. These mistakes have been observed in laboratory experiments, from the violations of the cancellation principle observed in the Allais' (1953) and Ellsberg's (1961) Paradoxes to the effects of framing, observed by Tversky and Kahneman (1981), among others.

Since our brains have to deal with these limitations, it is reasonable to assume that evolution would have provided us with heuristics that, although certainly not the same as full rationality, would be efficient given the limitations in our brain power and, for most problems of evolutive importance, capable of providing answers basically as good as possible to those of complete rationality. That is, humans probably use rules of thumb that allow them to make fast decisions, but that are subject to biases under some conditions. Simon (1951) proposed that we work with a bounded rationality, a concept that was later developed by a number of researchers, as reviewed in Gigerenzer and Selten (2001) or Selten (2001). Since any successful heuristics has to deal with the limits of our brains, they should be fast and frugal, as proposed by Gigerenzer and Goldstein (1996), but any rule of thumb we use also needs to be a good approximation to the results of complete rationality, at least for most of the problems our ancestors had to solve. Martignon (2001) has compared the decision-making heuristics with the optimal models of complete rationality, showing that there are circumstances where the fast heuristics can perform very well. Some models have shown that there are fast, surprisingly accurate heuristics for the problem of looking for the best action when there

are many different aspects to be considered and weighted in the decision process (see, per example the works of Hertwig (1999) and Selten (1998)).

On probabilistic biases, there have also been some results showing that there are efficient heuristics. Per example, Gigerenzer and Hoffrage (1995) have shown that people are capable of dealing better with frequencies than probabilities in a typical Bayesian problem, what makes sense since the problems dealt with by our ancestors would probably be related to the frequencies of observed events, certainly not with the mathematical laws of Probability. Kareev et al (1997) have noticed that the use of small samples can make the detection of correlation easier, providing an efficient heuristics for the problem of association between variables. However, there are still many other mistakes made in human judgment that have been left unexplained as good approximations to real problems, per example, the weighting functions (as per example, in Prospect Theory observed by Kahneman and Tversky, 1979), or the evaluation of compound events probabilities, as observed by Cohen et al. (1979), or the observation of conservatism on the update of opinions, made by Phillips and Edwards (1966).

As a consequence, there has been a recurrent critique of the heuristics program (Gilovich and Griffin (2002) have written a review about those critiques) based on arguments that say that human beings cannot be that dumb. Pinker (1997) has also proposed that any biases in human reasoning about probabilities should be due to the fact that our ancestors needed only a talent to work with frequencies and not real values of probabilities and it was indeed observed that humans have a tendency to work better with frequencies, as noted above. However, as we will see, simply accepting a statement about the observed frequency as the estimate for that frequency does not take into account several aspects of the complexity of inferential problems in the real world our ancestors had to deal with. In particular, this would completely ignore the fact that any

inference is subject to uncertainty, even more so those made by our ancestors from the limited number of observations they were available to collect for each problem.

We will see that the problem of the probabilistic biases can be understood as an adaptation, in an environment where there is uncertainty, errors, deception and a need to learn. In this sense, it is possible that our brains are actually correcting the probability values to values that would represent a better prediction in natural environments, but not necessarily in laboratory experiments or math classes. Special attention will be given to explaining the weighting functions w, that change the stated value p of the probability of an event to $w(p)$. Those functions are used in many theories describing human probability decisions, as per example, Prospect Theory (Kahneman and Tversky, 1979), Cumulative Prospect Theory, CPT (Kahneman and Tversky, 1992), as well as in models that describe paradoxes of CPT, such as transfer of exchange theory (Birnbaum and Chavez, 1997), or gains decomposition theory, by Luce (2000) and Marley and Luce (2001). All these models can be described as a general class of configural weight models, where the weight of a possibility can depend also on the other possibilities available. A good, recent review of these models, comparing their predictions with many experiments can be found in Birnbaum (2005).

We will see that the way we deal with probabilities, regardless of which descriptive theory is actually correct, might be closer to rational Bayesian inference than one would expect from the literature, where our decision skills are often described as erroneous. We will see that the laboratory findings are consistent with the hypothesis that our brains are built to work as if the stated values were actually obtained from observations of a small sample of results. Adaptive Probability Theory, APT, will be presented as the idea that our brains have a way to deal with probability that might have been efficient problem solving for real problems with limited data, although it does fail

when tested in laboratory with probabilities known exactly, something that wouldn't have happened to out ancestors. This article will show that APT can explain the observed S-shape of the weighting functions and we will see the ideas contained here are also able to explain the problems with coalescing and stochastic dominance violations that are in conflict with CPT. We will also see that the observed results that the weighting functions are not so well defined in the regions close to certainty can actually be seen as a simpler version of a sensitivity analysis. We will also propose possible explanations for the compound events probability problem and the conservatism biases.

The approach in this article will be neither normative nor descriptive, but an attempt to explain how the observed behavior concerning probabilities can be an approximation to the rules of decision making and an adaptation to solving the problems our ancestors had to deal with. In this sense, it should be noticed that I am not proposing that people do think exactly like the approximations bellow, only that APT provides reasonable approximations to rational decision making under the circumstances our ancestors lived. And, since APT is actually compatible with the observed laboratory biases, it is quite possible that this is the reason why we deal with probability the way we do. Whatever the real computations our brains do is a different problem, dealt with by the several descriptive theories available.

2. Detecting Relationships and Small Samples

If detecting true relations is important enough that detecting relations where there are none is an acceptable price to pay, a heuristics that allows for fast recognition of present correlations, as the one proposed by Kareev et al (1997), would be quite useful. Per example, if one has to choose between two possible actions, A and B, and

there is no a priori reason why one would lead to the best outcome with higher probability, simply flipping a coin is as useful as a device to make the decision as any other. However, if there is any covariate that has been observed before and this variable can help predicting the best course of action, it would be advantageous to detect it as soon as possible. In that sense, the observation that small samples actually can help on this detection provides a good example of a possible successful heuristics for dealing with probabilistic evaluations.

However, this early detection can certainly presents problems. It is easy to notice that this will lead to a number of wrong conclusions, especially false detections of correlation where none is present. Per example, if a sample with $n = 6$ observations for two uncorrelated variables is used, quite often, the sample will provide evidence that there is correlation and it won't be rare that this evidence will get strong. A simple simulation, where samples were obtained for two variables obeying a uniform distribution between 0 to 1 were generated independently and the observed correlation calculated. After 10,000 pairs of samples were generated, one can see the curve is higher near the correct value of 0, but it does not decrease very fast. As a matter of fact, almost 32% of the samples observed had a correlation (negative or positive) of 0.5 or higher. For samples not so small, per example, for 20 pair of points, that number drops reasonably fast to only about 2.5% and more than 60% of the observed correlations lie between -0.2 and +0.2. Kareev's results, however, are mostly important for the very small regions, for values of n under 8, showing that the early detection would work best if humans would actually use quite small samples. If, on one hand, using small samples help an early detection of correlates, without much trouble, on the other hand, using an uncorrelated variable to help with a decision process where one has no information available is not much different from flipping the coin, since it will basically

provide a random choice. If that is the best one could do, picking a wrong covariate will provide a result as good (or as bad) as anything available and increasing the chances of detecting a true correlate with less work, that is, with smaller costs, is actually a very good adaptation, even though it will lead to wrong decisions. It shouldn't be such a surprise, therefore, that people can be convinced by too weak evidence when no better alternatives are available.

3. Inference on Probabilities in Limited Samples

In the experiments that suggested that humans use weighting functions in their decision making process, the scientist would present people with the details of different bets they could choose from. Those bets differed by assigning different probabilities to different rewards and, usually, the probability value p_A (or simply p, as we will use from now on) that a certain outcome A would obtain was supposed to be known for certain. From a mathematical point of view, there was no uncertainty on the value of p, since the subsequent random draw would use exactly that value for chance and, therefore, our brain should use this information as a parameter known with certainty. However, in the environment where human minds developed, be it, from an evolutive point of view, the jungle where our ancestors lived, or, from a personal point of view, the environment where children are raised, certainty about the chances of an outcome is not something that will happen often, if at all. More likely, one has to infer, from a sample with finite size of previous observations, what is the probability that some event will happen again. In this sense, our minds would, in most daily situations, away from laboratory testing, arrive at much better results if this uncertainty was taken into account.

If the mind of a person is built to consider all probability evaluations as subject to error, as it is actually the case in real life problems, the information contained in p will be used to update her believes on the probability of outcome A, after some prior expectation about general probabilities. Therefore, it is reasonable to assume that the average probability will be a function of the stated probability, and that function can be called a weighting function $w(p)$. The correct way to update the prior knowledge is using the Bayes Theorem. For that, a model of how the information is obtained is necessary.

3.1 Estimating from a binomial likelihood

Of course, if there is no uncertainty, the correct inference is that the stated value of p is the real one. But that would never have happened to our ancestors. For them, all they had were observed frequencies and they'd have to deal with the uncertainty associated to that. If, in n observations, the outcome she was interested about was observed s times, that is, the observed frequency was s/n and if one takes into account the uncertainty, this becomes a classical problem of binomial inference on the probability of an event. The binomial likelihood is $l(s \mid p) \propto p^s (1-p)^{n-s}$ and, since we are looking for easy heuristics, one should choose the easiest approach, or the conjugate prior. Or, in other words, the prior should be a Beta distribution with parameters a and b (average equal to $a/(a+b)$), given by $f(p) \propto p^{a-1}(1-p)^{b-1}$ (for more details, check Bernardo, 1994). Since the posterior distribution is obtained, aside the normalization constant, by multiplying the likelihood and the prior, it will still be a Beta, with parameters $a+s$ and $b+n-s$, with an average value given by

$$\mu = \frac{a+s}{n+a+b} = \frac{a+pn}{n+a+b} \qquad (1)$$

No integration is actually necessary, if the mean is all one is interested about.

Notice that, if one only hears a probability statement, n is not known. Therefore, even in the simplest case, where the stated value p is supposed to be the observed frequency in a finite sample, if the sample size n is not known previously, it must be integrated out from a priori distribution for n, something that might be quite difficult to specify. If one knew the sample size, Equation 1 would provide the average estimate for the probability, that is, $w(p)$. In the previous section, we have seen that often humans use a small value for n, at least when looking for covariates. There is no reason to believe this would be different now. It seems reasonable to assume that n can not be too large, since our ancestors would, for most problems, not have a very large sample of observations to draw their conclusions from.

Therefore, we need some reasonable and solid assumptions about n. For a fixed sample size, Equation 1 provides a weighting function that is a straight line, but no longer the $w(p) = p$ line. However, for a fixed value of n, values of p close enough to 0 or 1 cannot be realistically achieved. Of course, it is also true that most values of p are not possible for any size of the sample, and this observation could be related with human inability to distinguish with close values of p. We have to keep in mind that what we are looking for is not a rigorous mathematical analysis, but a good heuristics that should approximate the rigorous analysis fairly well is some cases. Still, it is clear that, in order to observe a very unlikely event (or, equivalently, the non-occurrence of a very likely event), the size of n can not be small and must increase. An evaluation of a probability of 0.5 can be done (subject to error) with a small sample size, per example, 6, however the same sample size would never be able to predict a chance of 1 in 1,000. Therefore, we will assume that n is a function $n(p)$ of the probability p for a good

heuristics and it is also reasonable to assume that $n(p)$ diverges as $p \to 0$ or $p \to 1$. Given such a function, the value of $w(p)$ becomes determined uniquely for each p.

Of course, this is only a reasonable choice and not a demonstration. Still, values around 50% can mean simply that the person stating the value has no information at all about the problem, that is, there is a chance the sample size would be basically zero in that region. As the probability goes to one or zero, the person is actually saying she knows something about the problem and, at the very least, the chance that n is zero becomes smaller. In average, that means that our evaluation for n must increase as we go from the 50% towards the near certainty regions.

Figure 1 about here

Figure 1 shows the resulting weighting curves for the cases where $n \propto t^{-\gamma}$ and $n \propto \ln(t)$, where $t(p) = \min(p;1-p)$, compared to the curve proposed for the observed data by Prelec(2000), as well as the rational choice $w(p) = p$. The proportionality factor for the sample size dependence to p was chosen for all of the curves so that $n(0.5) = 4$, an arbitrary value that provided shapes close to those described in the literature. For larger sample sizes $n(0.5)$, the results started to become closer to the $w(p) = p$ curve too fast, as the data soon become stronger than the priori. This seems to indicate that our brains are working with the hypothesis that the sample sizes used for inference are rather small. The parameters for the priori distribution were taken to be $a = 1$ and $b = e - 1$, chosen as to provide the largest possible variance (given the constraints $a \geq 1$ and $b \geq 1$) compatible with the observation that the fixed point where $w(p) = p$ is actually close to $p = 1/e$, instead of $p = 0.5$. The uniform prior, representing no prior knowledge, would be given by $a = b = 1$ and this is the one we would expect to be the one used, since the problem is symmetric on the estimation of p

and $1-p$. We will return to the reasons why our brains might be working to approximate a different prior later. Of course, if this value was close to the average value of the probabilities humans had to estimate, this would be more than enough, but, as we will see, other reasons are also possible.

It should be noted that the case where $\gamma = 1$ corresponds to the case where the sample size increases fast enough so that an unlikely event is always expected to be observed, even for extreme values of p. This is not true for values of γ smaller than 1.0, although those values do provide a result closer to the observed curves. The fact that $n(p)$ does not increase fast enough for the curve observed in human reasoning can be understood as how much the data is considered trustworthy. Larger sample sizes mean the data is more important than the priori. Therefore, if there were some possibility of error or deception in the data, it would be a reasonable heuristics to assign it less strength by means of a smaller sample size and it seems reasonable that, if a too extreme probability was stated for our ancestors, that was probably due to exaggeration instead of observation. That is, it is reasonable to assume that $n(p)$ for an efficient heuristics might not increase as fast as it should, from a mathematical point of view.

3.2 Close to certainty behavior

Figure 2 about here

Another interesting feature to observe is the behavior of $w(p)$ as $p \to 0$, since that is the region where the sample size is not increasing fast enough. The behavior of the inferences as one gets close to certainty is shown in Figure 2. While the different curves agree reasonably well in the uncertainty regions, the ratio between the inferences obtained with different functions $n(p)$ and $m(p)$ will only go to 1.0 when p goes to zero if the both sample sizes $n(p)$ and $m(p)$ go to infinity at least as fast as curves

where γ is strictly above 1.0. Since this doesn't seem to be the case for our reasoning, that ratio will actually diverge as p goes to zero or to one, and the actual value of the inference becomes too sensitive to the choice of the prior model for $n(p)$. That is, by not been able to assign exact values for $w(p)$ for different values of probability (the difference between 1 in a million or 5 in a million depends on the context), our brains might be adopting an heuristics based on the fact that a sensibility analysis in this region will show that the values are approaching zero at very different speeds. Different priors will lead to different inferences, meaning the results obtained are not robust and, therefore, the value of $w(p)$ in this region is not so well defined in the sense of how fast it actually approaches zero.

4. Coalescing and Stochastic Dominance Violations

The results of the previous section agree with the observed behavior when a bet with only two possibilities is proposed to people. If three or more possibilities exist, new effects have been observed, as per example, violations of coalescing and stochastic dominance, that can be explained by some of the descriptive theories, but not all of them (in particular, CPT predicts that coalescing should be respected). As we will see, if one has to estimate probabilities with uncertainty, these properties can be explained in the same way we have explained the weighting functions in the previous section, by assuming our brains were built with heuristics that approximate rational behavior under certain circumstances.

4.1 Coalescing

Coalescing is the property that says it should be indifferent if a possible result with a certain probability and return is split into two, with the same return for both

possibilities and total probability equal to the original split probability. That is one should consider that the two following bets, A={$10,0.90; $100,0.10} and B={$10,0.90; $100,0.05; $100,0.05} (where each pair is a possible result, with the first number providing the return and the second, the probability of obtaining it), are equally desirable, since they are the same bet.

In order to test if people do obey coalescing, Birnbaum (2005) presents an experiment where each participant had to choose between two bets, A and B, and then made the choice between the bets A' and B', that were the same as A and B, except that they were presented with the alternatives that provided the same return coalesced, as we can see in Table 1.

Insert Table 1 about here

If a person respects coalescing, if she chose A over B, she should necessarily choose A' over B'. However, people showed a consistent tendency to choose B over A and A' over B' (almost half of the participants did switch from B to A'). Since this choice does contradict rationality, this seems to be another serious mistake in human reasoning.

However, this is not necessarily so. Notice that in one case we have two possibilities and in the other three possible outcomes. With three possible outcomes, if our brain is again considering, without our knowledge, that there is uncertainty in the probabilities, we need to obtain an approximation for the problem where we have to estimate p and q (the third will just be $1-p-q$). This is analogous to the binomial problem and it a simple extension of it will suffice. If one made n observations, from where x cases were observed to match the first possibility (probability p), y, the second and z, the last one, we would have a likelihood $l(s \mid p) \propto p^x q^y (1-p-q)^{n-x-y}$ and the easiest prior to use would be a extension of the Beta function to more variables,

that is, a Dirichlet function, given by $f(p,q) \propto p^{a-1}q^{b-1}(1-p-q)^{c-1}$. Notice that we have one extra parameter and now the uniform priori will be given by $a = b = c = 1$, that provides equal initial chances, 1/3, to each case. Notice that this is close to the average value we had to use to fit the curve in the last section, meaning that our brain might be using a priori close to the one proposed there because it was built to work with problems where there are not only two excluding possibilities, but three.

Once more, the expression for the average values for p and q have a very simple analytic expression, once the integration is done and are given by

$$w(p) = \frac{a + pn}{n + a + b + c}$$
$$w(q) = \frac{b + qn}{n + a + b + c} \qquad (2)$$

Supposing again that n is estimated from the most extreme probability, that is from $t = \min(p, q, 1-p-q)$, one can choose again $n \propto t^{-\gamma}$ and estimate the weighted probabilities for each of the prospects in Table 1. It can be thought that, since we have more possibilities, that, by itself should be enough to require a larger sample size and this is an issue that still needs to be explored further. For this example, it is assumed that only the probability values influence the sample size. One still has to choose a value for γ. Table 1 shows that values around 0.3 to 0.7 are good matches, but one can see from Table 2 that 0.3 provides a better match to Prelec function. Also, one has that, in this specific example, one must have γ less than 0.3966 if the observed results are to be explained.

If one takes $\gamma = 0.3$ (an arbitrary value compatible with the observations), the results for the weighting functions and the expected returns r can be seen in Table 2. In this case, the choices are actually B over A and A' over B', meaning that people will not obey coalescing. The priors were taken to be uniform in both cases.

Table 2 about here

At first sight, these results might seem strange, but they can be easily explained. First, since the second choice bets had one less possibility, the priors mean different things for each of them (they do mean the same if one decides to take $a = 1$ and $b = 2$). But, more important, since the sample size was chosen to depend on the extreme value, it is different in each case and it is actually smaller for the A' prospect, since it has less extreme probabilities.

4.2 Stochastic Dominance

Stochastic dominance means that one bet clearly dominates the other, providing results that are, at worst, equal to the second bet, the second should never be chosen over the first. Again, following an example of Birnbaum (2005), if one has G={$96,0.9; $12,0.1} and G+={$96,0.9; $14,0.05; $12,0.05}, it is easy to see that G+ clearly dominates G, since the only difference between the two is that the 10% branch of G bet was split into two possibilities, one paying the same amount and the other increasing the amount of money paid. However, if one repeats the analysis of the previous section, the extreme probability in G is 0.1 and in G+, 0.05, meaning that G+ will be analyzed as if resultant of a larger sample and, therefore, the weight function for the best unaltered result will be larger. The weighted expected returns are, this way, $r_G = 74.85$ and $r_{G+} = 79.68$. Again, our analysis of the situation indicates that, if there were uncertainty, G would not dominate G+ and G+ might be considered to provide a better expected return. For the worst bet proposed in Birnbaum, G-={$96,0.85; $90,0.05; $12,0.5}, where the largest possible outcome was split into two branches, one with lower return, one has a return of $r_{G-} = 81.48$, better not only than G, but also G+, despite the fact it is dominated by both other bets, agreeing once more with the observed violations of rationality.

5. Other Heuristics

APT can certainly be extended to other situations where humans seem to make probabilistic mistakes. Per example, Cohen et al. (1979) reported that people tend to overestimate the probability of conjunctive events. If people are asked to estimate the probability of a result in a two-stage lottery with equal probabilities in each state, their answer was far higher than the correct 25%, showing an average value of 45%.

This calculus is certainly wrong from a probabilistic point of view, where independence can be assumed and the value 0.5 is actually known for sure. However, if one was actually unsure of the real probability and only thought that, in average, the probability of a given outcome in the lottery was 50%, the independence becomes conditional on the value of p. The chance that two equal outcomes will obtain is given by p^2 and, since p is actually unknown, one has to integrate it in order to have an average estimate for that, getting, for a uniform priori, $\int_0^1 1.p^2 dp = 1/3$.

An alternative way to understand it, is to notice that, if one considers the two draws as one at a time, instead of both results, the first result should be used as inference about the real probability and we'd have, from assuming initially an uniform priori (50% chance), that the probability of the same result happening the next time would actually be, in average, 2/3, meaning that the compound probability of two equal results would be altered to 1/3. That is, for real problems where only conditional independence exists, the result is not the correct 25% for the situation where p is known to be 0.5 with certainty. Of course, if the uncertainty in the priori was smaller, the result would become closer to 25%.

Furthermore, if the conditional independence hypothesis is also dropped, the predicted results can become even closer to the actual observed behavior. And in many situations, especially when learning about some system where not much is actually known, even conditional independence might become a too strong assumption. Suppose, per example, that our ancestors wanted to evaluate the probability of finding predators at the river they used to get water from. If a rational man had a prior uniform distribution for the chance the predator would be there and, after that, only one observation made an hour ago where the predator was actually seen, the average chance a predator would be by the river would be increased, as above, to 2/3. However, if he wanted to go to the river only an hour later again, the events would not be really conditionally independent, as the predator might still be there. The existence of correlation between the observations implies that the fact the predator had been spotted earlier should make it more likely to observe it there again, increasing the probability from 2/3. That is, the observed estimate around 45% can be at least partially explained as someone trying to make inferences when the concepts of independence or even conditional independence do not necessarily hold.

It is important to keep in mind that our ancestors had to deal with a world they didn't know how to describe and model as well as we do nowadays. It would make sense for a successful heuristics to include the learning about the system in study and, in that sense, the notion of independent sampling for similar events might not be natural in every case, as we have just seen. When faced with the same situation, not only the previous result can be used as inference for the next ones, but also it might have happened that some covariance between the results existed and this might be the origin of the conjunctive events bias.

Another observed bias that can also be explained as a heuristics adapted to estimating probabilities in real, complex problems is conservatism. Conservatism can be defined as the fact that people seem to update their probability estimates slower than the rules of probability dictate, when presented with new information. That is, given their prior estimates and new data, the data is given less importance than it should and the subjective probabilities, after learning the new information, changes less than it should. This heuristics can be at least partially explained if one includes in the description of the problem the possibility that the new information can be erroneous or deceptive. If the probability of deception, including here also errors, is sufficiently large, the value of new data should be challenged and the estimates should actually change slower than the simpler calculation, not including this possibility, would show. This, of course, does not mean that people actually distrust the reported results, at least, not in a conscious way. Instead, it is the heuristics that work inside our brains that might have evolved in a world where the information available was subject to all kind of errors. Notice that the same effect seems to be happening for the assumed sample sizes for extreme probabilities, since it does not grow as fast as it should.

6. Conclusions

We have seen that a number of probabilistic heuristics used by people can be explained by APT, that is, the idea that our minds might be adapted to a world far more complex than that of laboratory experiments or Probability classes and that our brains somehow are built to deal with probabilities that are uncertain. These heuristics were shown to be actually close to the results of Bayesian inferences, under the right circumstances, despite the fact that, in laboratory situations, where the probabilities are supposed to be exact, they do provide wrong answers. The weighting functions (as in

CPT or in configural weight models) can be understood as equivalent to a problem where the person who listens to a stated probability value believes that this value is actually an observed frequency, obtained from a finite and small sample, as long as some reasonable assumptions about the sample size as a function of the probability are made. It was shown that, since the heuristics is very sensitive to the prior for the sample size when one gets close to certainty, it is natural that the weighting functions will be uncertain in those regions. Since a small sample size can a good solution when obtaining information has a cost, it is reasonable that our mental processes would assume that stated probabilities were actually observed frequencies subject to sampling error and try to correct this. However, it should be clear that the calculations done above are not necessarily those performed by our brains, they only show that our biases are actually a better approximation than previously thought.

We have also seen that the observed violations to CPT can also be explained and that APT can explain why people violate coalescing and stochastic dominance. Although both rules are good normative rules, the laboratory tests assume that probabilities are known with certainty, an assumption our brains seem not to be built to work with. Under uncertainty, the sample size becomes an important factor and splitting probabilities can change the evaluation of the sample size used, therefore affecting all reasoning. If our brains are built to approximate these results, they actually should violate those principles, although it is still true that the laboratory observations are mistakes in our reasoning. Evolution does not provide the best possible solution, only one that works well in the environment where our ancestors lived.

The application of the idea that what human mind does is to analyze problems as an inference problems subject to errors, deception, and where independence is not assumed, was also capable of providing a possible explanation on the conjunctive

events as well as the conservatism biases, making it clear that the probabilistic biases in human decision-making is completely compatible with the evolutionary point of view. The laboratory observations and the descriptive theories can be better explained as a result of adaptations to a complex world. All the heuristics seem to work with the supposition that the world is actually more complex than the one tested in laboratory experiments, that is our brains is probably better adapted to real life than to laboratory problems, as one should expect. In this sense, these heuristics might have provided our ancestors with an efficient brain, capable of evaluating chances in a competent way, but not necessarily capable of performing correct probability calculations.

Adaptive Probability theory is a way to explain why we make the errors we make, showing they are actually good heuristics for real problems. It is neither a completely descriptive theory nor normative, since it uses the prescriptions of rational behavior to obtain reasonable approximations that we have seen match the observed behavior. In that sense, it is a theory to answer not how we do, but why we do it. More complete comparisons with the existent descriptive models are currently being prepared and tests to check how close APT actually described human behavior are being planned.

Acknowledgements

This work was partially done while the author was at the Faculdade de Ciências e Letras de Assis da Universidade Estadual Paulista. The author would also like to thank José Roberto Securato for the support during the preparation of part of this work and Jerome R. Busemeyer for pointing out some recent works that provided extra support to the ideas discussed here.

References

Allais, P. M. (1953) The behavior of rational man in risky situations - A critique of the axioms and postulates of the American School. *Econometrica*, **21**, 503-546.

Bernardo, J. M., Smith, A.F.M. (1994), *Bayesian Theory*. New York, Wiley.

Birnbaum, M. H., (2005) New paradoxes of risky decision making. Working Paper.

Birnbaum, M. H., & Chavez, A. (1997) Tests of theories of decision making: Violations of branch independence and distribution independence. *Organizational Behavior and Human Decision Processes*, **71** (2), 161-194.

Cohen, J., Chesnick, E.I., & Haran, D., (1979) Evaluation of compound probabilities in sequential choice, *Nature*, **232**, 414-416

Ellsberg, D. (1961) Risk, ambiguity and the Savage axioms. *Quart. J. of Economics*, **75**, 643-669.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psych. Rev.*, **103**, 650-669.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psych. Rev.*, **102**, 684-704.

Gigerenzer, G., & Selten, R. (2001). Rethinking Rationality i*n* G. Gigerenzer, R. Selten (eds.), *Bounded rationality: The adaptive toolbox. Dahlem Workshop Report*, 147-171. Cambridge, Mass, MIT Press.

Gilovich, T., & Griffin, D. (2002) Introduction - Heuristics and biases: Then and now *in* Gilovich, T., Griffin, D., & Kahneman, D. (eds.) *Heuristics and biases: The Psychology of intuitive judgment*. Cambridge, Cambridge University Press.

Hertwig, R., Hoffrage, U., & Martignon, L. (1999). Quick estimation: Letting the environment do the work *in* G. Gigerenzer, & P. M. Todd (eds.), *Simple Heuristics that make us smart*, 209-234. New York, Oxford University Press.

Kahneman, D., & Tversky, A. (1979) Prospect theory: An analysis of decision under risk. *Econometrica*, **47**, 263-291

Kahneman, D., & Tversky, A. (1992) Advances in prospect theory: Cumulative representation of uncertainty. *J. of Risk and Uncertainty*, **5**, 297-324.

Kareev, Y., Lieberman, I., & Lev, M. (1997) Through a narrow window: Sample size and the perception of correlation. *J. of Exp. Psych.: General*, **126**, 278-287

Luce, R. D. (2000) *Utility of gains and losses: Measurement-theoretical and experimental approaches.* Mahwah: Lawrence Erlbaum Associates.

Marley, A. A. J., & *Luce*, E. D. (2001) Rank-weighted utilities and qualitative convolution. *J. of Risk and Uncertainty*, **23** (2), 135-163.

Martignon, L. (*2001*). Comparing fast and frugal heuristics and optimal models *in* G. Gigerenzer, & R. Selten (eds.), *Bounded rationality: The adaptive toolbox. Dahlem Workshop Report*, 147-171. Cambridge, Mass, MIT Press.

Phillips, L. D., & Edwards, W. (1966) Conservatism in a simple probability inference task, *J. of Exp. Psych.*, **72**, 346-354.

Pinker, S. (1997) *How* the *mind works*. New York, Norton.

Prelec, D. (2000) Compound invariant weighting functions in Prospect Theory *in* Kahneman, D., & Tversky, A. (eds.), *Choices, Values and Frames*}, 67-92. New York, Russell Sage Foundation, Cambridge University Press.

Savage, L. J. (1954) *The Foundations of Statistics*. New York, Wiley.

Selten, R. (1998), Aspiration adaptation theory. *J. of Math. Psych.*, **42**, 191-214.

Selten, R. (2001). What is bounded rationality? *in* G. Gigerenzer, R. Selten (eds.), *Bounded rationality: The adaptive toolbox. Dahlem Workshop Report,* 147-171. Cambridge, Mass, MIT Press.

Simon, H. A. (1956) Rational choice and the structure of environments. *Psych. Rev.*, **63**, 129-138

Tversky, A., & Kahneman, D. (1981) The framing of decisions and psychology of choice. *Science*, **211**, 453-458

von Neumann, J., & Morgenstern, O. (1947) *Theory of Games and Economic Behavior*. Princeton, Princeton University Press.

A: 85% to win \$100 10% to win \$50 5% to win \$ 50	B: 85% to win \$100 10% to win \$100 5% to win 7
A': 85% to win \$100 15% to win \$50	B': 95% to win \$100 5% to win 7

Table 1: Bets used by Birnbaum (2005) to test coalescing. Instead of presenting the bets as frequencies, as in the test reported, they are presented here as probabilities, for ease of comparison with the equations.

A: $w(0.85) = 0.709$ $w(0.10) = 0.164$ and $r = 85.4$ $w(0.05) = 0.127$	B: $w(0.85) = 0.709$ $w(0.10) = 0.164$ and $r = 88.2$ $w(0.05) = 0.127$
A': $w(0.85) = 0.76$ $w(0.15) = 0.24$ and $r = 88.0$	B': $w(0.95) = 0.86$ $w(0.05) = 0.14$ and $r = 87.0$

Table 2: Weighted probabilities and expected returns for the bets for $\gamma = 0.3$ and uniform priors.

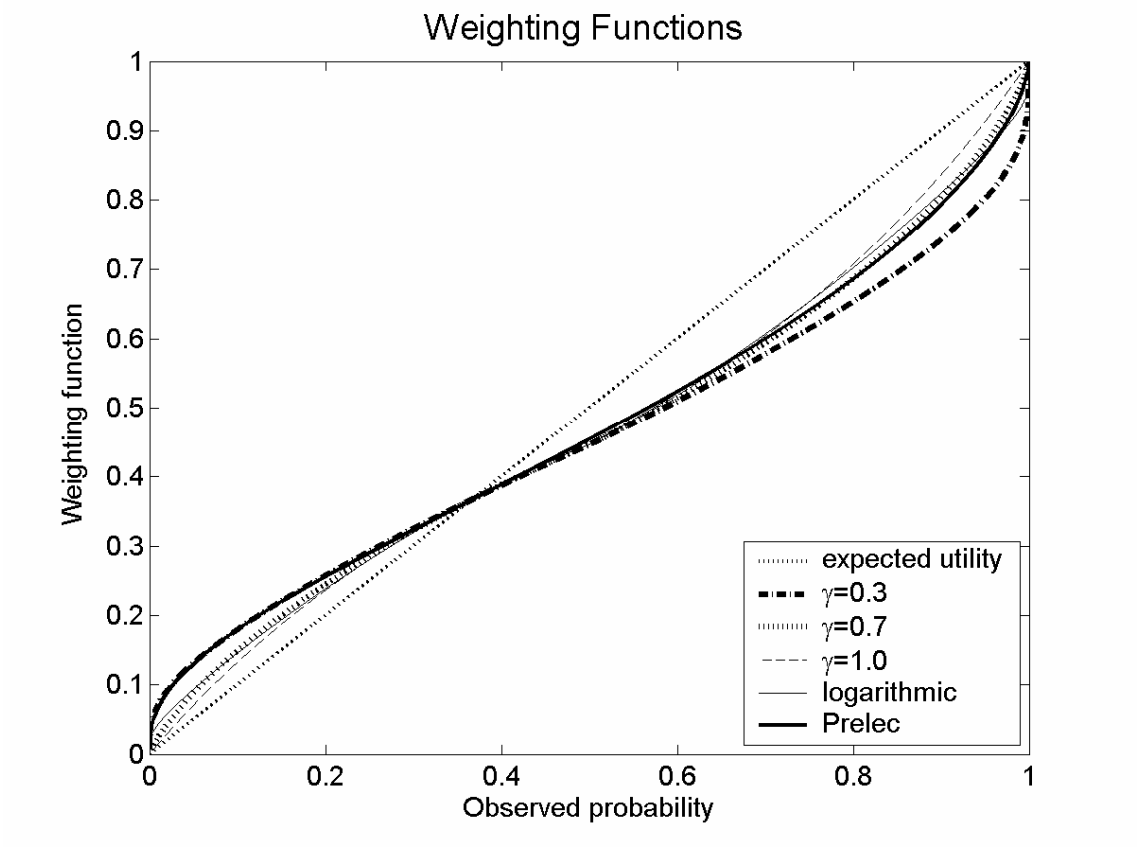

Figure 1: Weighting functions as a function of the stated probability, for the expected utility case compared to the logarithmic curve as well as the curves for a few values of γ. The functional shape proposed by Prelec(2000) is also shown for comparison.

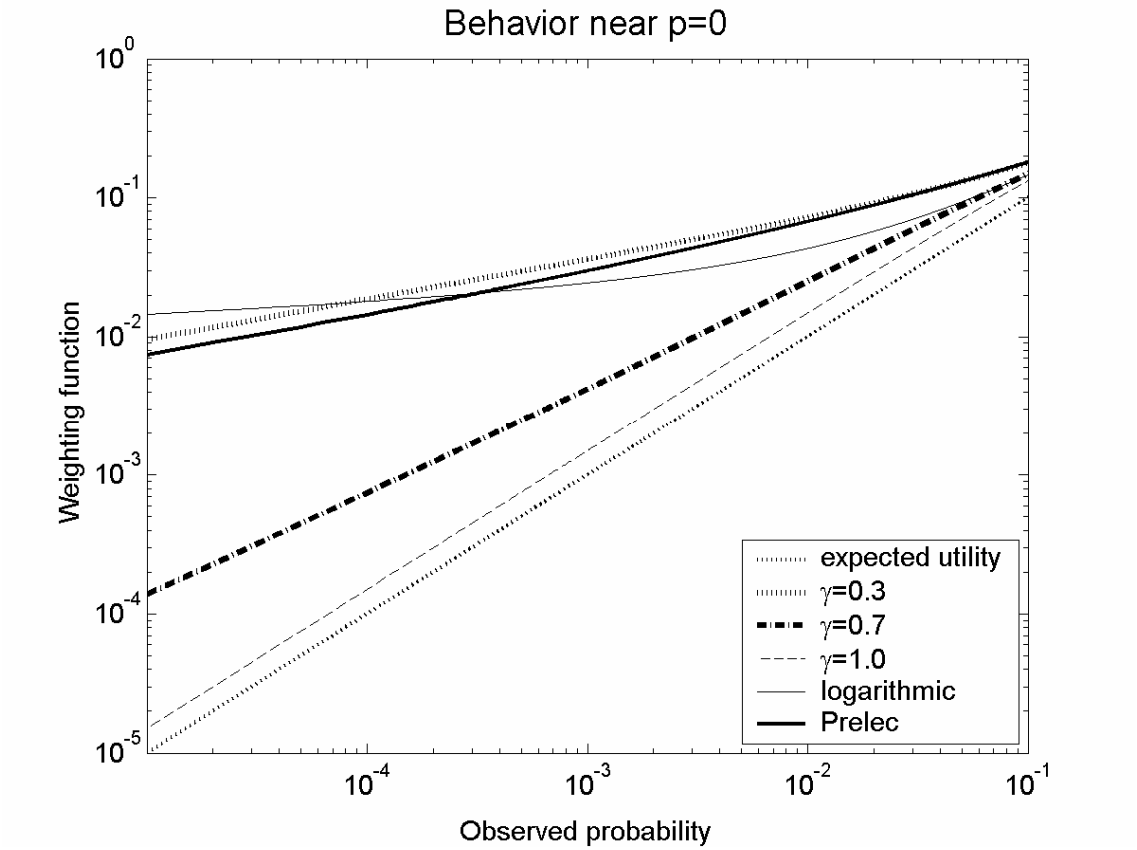

Figure 2: Behavior of the weighting functions as $p \to 0$.