What Comes to Mind

Nicola Gennaioli and Andrei Shleifer[1]

Second Draft, February 13, 2009

Abstract

We present a model of judgement under uncertainty, in which an agent combines data received from the external world with information retrieved from memory to evaluate a hypothesis. We focus on what comes to mind immediately, as the agent makes quick, System I, evaluations. Because the automatic retrieval of data from memory is both limited and selected, the agent's evaluations may be severely biased. Some of the heuristics and biases evidence presented by Kahneman and Tversky, including conjunction and disjunction fallacies, can be accounted for in this framework.

[1] CREI and Harvard University, respectively. We are deeply grateful to Josh Schwarzstein for considerable input, and to Xavier Gabaix, Elizabeth Kensinger, Scott Kominers, David Laibson, Sendhil Mullainathan, Giacomo Ponzetto, Drazen Prelec, Antonio Rangel, Jesse Shapiro, Jeremy Stein, and Richard Thaler for extremely helpful comments. Gennaioli thanks the Spanish Ministerio de Ciencia y Tecnologia and the Barcelona Graduate School of Economics for financial support.

1. Introduction

Since the early 1970s, Daniel Kahneman and Amos Tversky (hereafter KT 1972, 1974, 1983, 2002) published a series of remarkable experiments documenting significant deviations from Bayesian theory of judgment under uncertainty. While KT's heuristics and biases program has survived substantial experimental scrutiny, models of heuristics have proved elusive[2]. In this paper, we offer a new model of decision making that accounts for quite a bit of this experimental evidence.

Our approach is succinctly captured by an observation made by Kahneman in a 2008 lecture at Harvard. Kahneman noted that heuristics describe how people evaluate hypotheses quickly, based on what first comes to mind. People may be entirely capable of more careful deliberation and analysis, and perhaps of better decisions, but not when they do not think things through. Kahneman (2003) describes such quick decision making as System 1 (intuition), and distinguishes it from System 2 (reasoning). We present a formal model of such System 1 judgement, based on what comes to mind.

We describe a problem in which a decision maker evaluates a hypothesis in light of some data, but with some residual uncertainty remaining. This residual uncertainty can be thought of as scenarios that have not been specified. We think of the decision maker as automatically filling in from memory some of the scenarios, but not others, and making the judgement in light of what he is thinking about. Our approach is broadly consistent with KT's insistence that judgment under uncertainty is similar to perception. Just as an individual fills in details from memory when interpreting sensory data (for example, when looking at the duck-rabbit or when judging distance from the height of the

[2] Partial exceptions include Mullainathan (2000), Griffin and Tversky (1992), and Tversky and Koehler (1994), to which we return.

object), the decision maker recalls missing scenarios when he evaluates a hypothesis.

In our model, what is automatically retrieved from memory – what comes to mind – in the first instance is both *limited* and *selected.* On the one hand, some scenarios come to mind immediately, others do not: the working memory is limited. On the other hand, the selection is primed by the question being asked (or hypothesis being evaluated), and might not be the data a Bayesian would ask for. Crucially, we specify that scenarios come to mind in order of their diagnosticity, which formally means their ability to predict the hypothesis being evaluated relative to other hypotheses. Diagnosticity captures the idea that we recall more easily "representative" scenarios for the hypothesis evaluated by the agent. In this model, when the decision maker only thinks of some scenarios, his evaluations could (but need not) be severely biased; if he considers all the scenarios, his decisions are rational in the Bayesian sense. The deliberate System 2 evaluations thus emerge as the limiting case of System 1 judgments, as more things come to mind.

In the next section, we present a simple example illustrating our approach. In Section 3, we present the formal model, and discuss in detail the relationship of our approach to prior work. The following sections apply the model to KT's experimental findings. Section 4 considers some of biases related to representativeness, such as base rate neglect and insensitivity to predictability. Section 5 addresses the failures of extensionality, namely the conjunction and disjunction fallacies. Section 6 concludes.

2 An Example: Electoral Campaign

We illustrate the basic working of our model using an example loosely based on Popkin (1991). An Asian-American voter evaluates the qualifications of a presidential candidate after the latter fails to use chopsticks to eat noodles at a campaign banquet. The voter classifies candidates along two dimensions: qualification and familiarity with the Asian customs and community. The voter estimates the probability that the candidate is qualified, which is all he cares about, but along the way fills in the candidate's familiarity with Asian customs, which we call "scenarios." Think of the voter as having a database of "associations" in his long term memory, summarized by a distribution of candidate types that, conditional on failing to use chopsticks, is described in Table 1.A:

Candidate cannot use chopsticks		Familiarity with Asian customs	
		familiar	*unfamiliar*
qualification of candidate	*qualified*	0.024	0.43
	unqualified	0.026	0.52

Table 1.A

Table 1.A captures two ideas: i) failure to use chopsticks is very informative about unfamiliarity with Asian customs (95% of the candidates who fail to use chopsticks are "unfamiliar" with Asian customs), but ii) familiarity with Asian customs is scarcely informative about qualification (relative to a prior of 1/2). The latter property is reflected in the qualification estimate of a Bayesian voter, which is equal to:

$$\Pr(\textit{qualified}) = \Pr(\textit{qualified}, \textit{familiar}) + \Pr(\textit{qualified}, \textit{unfamiliar}) = 0.454 \qquad (1)$$

The Bayesian reduces his prior very little due to the event's low informational content.

Although Table 1.A is stored in the voter's long term memory, due to working memory limits not all candidate types come to his mind to aid his evaluation of the candidate's qualification. In equation (1), the Bayesian voter considers that both

qualified and unqualified candidates can be either familiar or unfamiliar with Asian customs. The decision maker we describe, in contrast, is a "local thinker," named so because, to evaluate hypotheses, he does not use all the data in Table 1.A but only the information he obtains by sampling in his memory some specific examples of qualified and unqualified candidates. In KT's spirit, what first comes to the agent's mind are examples of *representative*, or stereotypical, qualified and unqualified candidates.

We model this idea by assuming that the voter draws from memory examples of qualified and unqualified candidates by searching for the most diagnostic levels of familiarity – scenarios – for each type. These scenarios are respectively given by:

$$s(qualified) = \underset{s \in \{familiar, unfamiliar\}}{\arg\max} \Pr(qualified|s), \quad (2)$$

$$s(unqualified) = \underset{s \in \{familiar, unfamiliar\}}{\arg\max} \Pr(unqualified|s). \quad (3)$$

Each type of candidate brings to mind examples sharing the level of familiarity relatively more associated with that type. In Table 1.A, this means that a qualified candidate evokes examples of candidates that are familiar with Asian customs, but an unqualified candidate evokes candidates unfamiliar with them.[3] This is because, for this voter, a candidate familiar with his own customs is at least marginally more qualified, so qualification and familiarity are associated in the stereotypical qualified candidate. This effectively reduces the voter's information to the circled diagonal below:

Candidate cannot use chopsticks		Familiarity with Asian customs	
		familiar	*Unfamiliar*
qualification of candidate	*qualified*	0.024	0.43
	unqualified	0.026	0.52

Table 1.B

[3] Indeed, Pr(*qual*|*fam*)=12/25>43/95=Pr(*qual*|*unfam*). The reverse is true for an unqualified candidate.

As the local thinker retrieves from memory stereotypical qualified and unqualified candidates, his assessment (indicated by superscript L) is only based on these stereotypes.

$$\Pr^{L}(qualified) = \frac{\Pr(qualified, familiar)}{\Pr(qualified, familiar) + \Pr(unqualified, unfamiliar)} \approx 0.044 \qquad (4)$$

Relative to a Bayesian, the local thinker overreacts to the candidate's inability to use chopsticks, underestimating his qualification by a factor of about 20! Why is this so?

Intuitively, inability to use chopsticks evokes in the voter's mind many examples of unqualified and unfamiliar candidates and few examples of qualified and familiar ones. The reason is that many stereotypical unqualified candidates are indeed unable to use chopsticks while most stereotypical qualified candidate can use them, which causes under-sampling and thus under-estimation of qualification. Under-estimation here is severe because, by recalling stereotypes, the voter forgets that many qualified candidates are unfamiliar with Asian customs! As noted by Popkin (1991), voters fit political facts using candidates' personal data because those data, even if uninformative, allow voters to map the candidate into a representative candidate. In our example, this effect is due to imperfect recall and causes a drastic bias in the estimate of qualification.

The same idea, though, suggests that in many cases local thinkers can produce fairly good assessments. Suppose for instance that the candidate claims that the main meat eaten by the Chinese is dog. Suppose that the distribution of candidate types is:

The Chinese only eat dogs		Familiarity with Asian customs	
		Familiar	*unfamiliar*
Qualification of candidate	*qualified*	0.25	0.025
	unqualified	0.025	0.7

Table 1.C

Table 1.C captures two ideas: i) the candidate's statement is quite informative about his unfamiliarity with the Asian customs and ii) unfamiliarity in this case is extremely informative about the candidate's qualification. Based on Table 1.C, a Bayesian assesses Pr(*qualified*) = 0.275. Due to ii) the local thinker still associates familiarity with qualification, and estimates:

$$\Pr^{L}(qualified) = \frac{\Pr(qualified, familiar)}{\Pr(qualified, familiar) + \Pr(unqualified, unfamiliar)} \approx 0.26 \qquad (5)$$

which is almost identical to a Bayesian's assessment. In contrast to the previous case, the candidate's unfamiliarity with the Asian customs is now so grotesque as to be very informative about the candidate's low qualification. In this case, local thinking generates a very mild loss of information.

Why this difference in the examples? After all, in both examples the stereotypical qualified (resp. unqualified) candidate is someone familiar (resp. unfamiliar) with Asian customs. In the first example of Table 1.B, though, the stereotypical qualified candidate is extremely uncommon because the bulk of qualified candidates are unfamiliar [i.e. a fraction 0.43/(0.024+0.43) = 0.95 of them], causing gross under-estimation of qualification. In contrast, in the second example, almost all of the qualified candidates are familiar with Asian customs. In this latter case, the stereotype of a qualified candidate is not only diagnostic but also likely, which greatly reduces the agent's assessment bias. As we show below, the relationship between the diagnosticity and likelihood of stereotypes is a key determinant of the accuracy of a local thinker's probabilistic assessments.

We now formalize our model of decision making and study its broader implications for judgment under uncertainty and the heuristics and biases research.

3 The Model

The world is described by a probability space (X, π), where $X \equiv \prod_{i=1,...,K} X_i$ is a finite state space generated by the product of $K \geq 1$ dimensions and the function $\pi : X \to [0,1]$ that maps each element $x \in X$ into a probability $\pi(x) \geq 0$ such that $\sum_{x \in X} \pi(x) = 1$. In the example of Section 2, the dimensions of X are the candidate's qualification and familiarity with Asian customs (i.e., $K = 2$), the elements $x \in X$ are candidate types and the probability space (X, π) is described in Table 1.A.

An agent evaluates the probability of $N > 1$ hypotheses $h_1,...,h_N$ in light of data d. Hypotheses and data are events of X; that is, $h_r, d \subseteq X$ for every $r = 1,..., N$. If the agent receives no data, then $d = X$: nothing is ruled out. Hypotheses may be non-exclusive and non-exhaustive. In (X, π), the probability of h_r is given by the formula:

$$\Pr(h_r | d) = \frac{\Pr(h_r \cap d)}{\Pr(d)} = \frac{\sum_{x \in h_r \cap d} \pi(x)}{\sum_{x \in d} \pi(x)}, \tag{6}$$

which integrates the probabilities of all elements consistent with the hypothesis *and* the data [i.e. $x \in h_r \cap d$], dividing the resulting sum by the probability of d alone. In our example, expression (1) follows from (6) since in Table 1.A the probabilities are normalized by Pr(*no chopsticks*). As we saw in Section 2, a local thinker may fail to produce the correct assessment (6) because he only considers a subset of elements x, those belonging to what we henceforth call his "represented state space".

3.1 The Represented State Space

The represented state space is shaped by the recall of elements in X prompted by

the assessed hypotheses h_r, $r = 1,\ldots,N$. Recall is governed by two assumptions. First, working memory limits the number of elements recalled by the agent to represent each hypothesis. Second, the elements recalled to represent a hypothesis are the most "diagnostic" ones for that hypothesis. Note that an element here is what we called a stereotype in the example of Section 2. We formalize the first assumption as follows:

A1 (Local Thinking): Given data d, the agent represents hypothesis h_r, $r = 1,\ldots,N$ by using at most $b \geq 1$ elements $x \in h_r \cap d$.

The set $h_r \cap d$ is the set of representations of hypothesis h_r and includes all the elements in X consistent with hypothesis h_r and with the data d. Two polar cases are of interest: i) the case where $b = 1$ when thinking is fully local and only one element in the set of representations is selected for each hypothesis, and ii) the case where b is sufficiently large that all hypotheses are represented using all elements in $h_r \cap d$. In the latter case, we say that the agent's representation of all hypotheses is perfect.[4]

The representation of hypothesis h_r is perfect if there are fewer than b elements in the set of representations $h_r \cap d$. At the extreme, if the hypothesis and the data identify a single element in X, even the representation by the agent with $b = 1$ is perfect. The more interesting case involves broad hypotheses consisting of more than b elements. In this case, when $b = 1$ the set of possible representations $h_r \cap d$ must be collapsed into a single element. To do so, the agent must attribute exact values to the dimensions of X that are not pinned down by the hypothesis and the data. For instance, in the example of Section 2, to represent qualified and unqualified candidates, the voter attributes one level

[4] A.1 is one way to capture limited recall. None of our substantive results would change if we alternatively assumed that the agent discounts the probability of certain elements.

of familiarity to each hypothetical level of qualification. We call such fitted levels of familiarity "scenarios".

To give a general definition of scenarios, suppose that h_r and d specify exact values (rather than ranges) for some dimensions of X, taking the form:

$$\left\{x \in X \middle| x_i = \hat{x}_i \quad for \quad some \quad \hat{x}_i \in X_i \quad and \quad some \quad i \in [1,...,K]\right\}, \tag{7}$$

where x_i denotes the i th dimension of element $x \in X$, while $\hat{x}_i$ is the exact value taken by such dimension in the hypothesis or data. The remaining dimensions are unrestricted. This is consistent with the example of Section 2 where hypotheses specify a qualification level, data specifies inability to use chopsticks, and the remaining familiarity dimension is left completely free. The possible scenarios for hypothesis h_r are defined as follows:

Definition 1. Suppose that $h_r \cap d$ fixes the values of $N_r < K$ dimensions in X. Denote by S the set of the remaining $K - N_r$ free dimensions. A scenario s is any event $s \equiv \left\{x \in X \middle| x_t = x_t' \quad for \quad some \quad x_t' \in X_t \quad and \quad all \quad t \in S\right\}$. If $N_r = K$, a scenario is $s = X$.

A scenario completes the details missing from the hypothesis and data because it identifies a single element in $h_r \cap d$: $s \cap h_r \cap d \in X$. A scenario can be viewed as a "frame", namely as a mental model allowing one to interpret a situation in light of partial data. But how do scenarios come to mind? We assume that the agent represents hypotheses taking the form of expression (7) in the following way:

A2 (Recall by Diagnosticity): Fix d and h_r. When $b = 1$, the agent represents h_r with the most "diagnostic" scenario s_r^1, which is the scenario maximizing:

$$\Pr(h_r \middle| s \cap d) = \frac{\Pr(h_r \cap s \cap d)}{\Pr(h_r \cap s \cap d) + \Pr(\overline{h_r} \cap s \cap d)}, \tag{8}$$

where $\overline{h_r}$ is the complement X/h_r in X of hypothesis h_r. When $b > 1$, the agent represents h_r with b most "diagnostic" scenarios s_r^k, $k = 1,\ldots,b$, where scenarios with a lower index k obtain a higher value of (8).

The local thinker represents h_r by recalling only the b most "diagnostic" scenarios, those that are more associated with h_r relative to the other hypotheses. Scenario s_r^1 can be interpreted as the most representative model for the hypothesis h_r because, together with the data, it maximizes the likelihood $\Pr(h_r|s \cap d)$ of the hypothesis.

We can derive the represented state space from the recalled scenarios. If the agent recalls s_r^1 in conjunction with the hypothesis h_r, he includes the corresponding element $s_r^1 \cap h_r \cap d \in X$ in the representation of h_r. Applying this logic to all the hypotheses $h_1,\ldots,h_N$ evaluated by the agent yields:

Definition 2 Denote by $M_r > 1$ the total number of scenarios for hypothesis h_r, $r = 1,\ldots,N$. Then, the agent's represented state space is $\bigcup_{\substack{r=1,\ldots,N \\ k=1,\ldots,\min(M_r,b)}} s_r^k \cap h_r \cap d$.

The represented state space is simply the union of all elements recalled by the agent for each of the assessed hypotheses. Definition 2 applies to hypotheses of the form given by (7), but it is easy to extend it to hypotheses which, rather than attributing exact values, restrict the range of some dimensions of X. Appendix 1 shows how to do that and to apply our model to the evaluation of these hypotheses as well.

3.2 Probabilistic Assessments by a Local Thinker

In the represented state space, the local thinker computes the probability of h_t as:

$$\Pr^L(h_t|d) = \frac{\sum_{k=1}^{\min(M_t,b)} \Pr(s_t^k \cap h_t \cap d)}{\sum_{r=1}^{N} \sum_{k=1}^{\min(M_t,b)} \Pr(s_r^k \cap h_r \cap d)}, \tag{9}$$

namely as the probability of the *representation* of h_t divided by that of the *representation* of all hypotheses h_r, $r = 1,\ldots,N$. One property of (9) is that the assessed probability of a hypothesis depends on the other hypotheses examined in conjunction with it. This is one key way in which the examined hypotheses shape assessments in our model. Evaluated at $b = 1$, (9) is the counterpart of expression (4) in Section 2.

We can rewrite (9) as:

$$\Pr^L(h_t|d) = \frac{\left[\sum_{k=1}^{\min(M_t,b)} \Pr(s_t^k|h_t \cap d)\right] \Pr(h_t \cap d)}{\sum_{r=1}^{N} \left[\sum_{k=1}^{\min(M_r,b)} \Pr(s_r^k|h_r \cap d)\right] \Pr(h_r \cap d)}, \tag{9'}$$

Suppose that the hypotheses examined are exhaustive, that is $\sum_{r=1}^{N} \Pr(h_r \cap d) = \Pr(d)$.

Expression (9') highlights the role of local thinking. If $b \geq M_r$ for all $r = 1,..,N$, then the bracketed terms disappear because $\sum_{k=1}^{M_r} \Pr(s_r^k|h_r \cap d) = 1$ must necessarily hold. In this case (9') boils down to $\Pr(h_t \cap d)/\Pr(d)$, which is the Bayesian's estimate of $\Pr(h_t|d)$. Biases in judgement can only arise when the agent's representations are limited, that is, when $b < M_r$ for some r.

To interpret (9'), note that $\Pr(s|h_r \cap d)$ is the likelihood of scenario s for h_r, of the probability of s when h_r is true. The bracketed terms in (9') then measure the share of a hypothesis' total probability captured by its representation. Equation (9') says that if the representations of all hypotheses are equally likely (bracketed terms are equal), the

estimate is perfect, even if memory limitations are severe. Otherwise, biases may arise. Despite the importance of likelihood for the accuracy of assessments, the ranking of scenarios by their likelihood often differs from that by their diagnosticity.

3.3 Discussion of Setup and Assumptions

It is worth discussing the conceptual structure of the model. Assumption A2 posits that a hypothesis is represented using a mental model, or more specifically a scenario, that is most closely associated with this hypothesis relative to other ones, much in the spirit of KT's notion of *representativeness*. Representativeness is "defined as a subjective judgement of the extent to which the event in question is similar in essential properties to its parent population or reflects the salient features of the process by which it is generated" (KT 1972, p 431). Indeed, KT (2002, p.23) have a discussion of diagnosticity related to our model's definition: "Representativeness tends to covary with frequency: common instances and frequent events are generally more representative than unusual instances and rare events," but they add that "an attribute is representative of a class if it is very diagnostic; that is the relative frequency of this attribute is much higher in that class than in a relevant reference class." In other words, sometimes what is representative is not likely. As we show below, the representation of hypotheses by diagnostic but unlikely scenarios drives many of the KT anomalies.

Our approach is related to Griffin and Tversky's (1992) notion that agents assess a hypothesis more in light of the *strength* of the evidence in its favour, a concept akin to our "diagnosticity", rather than in light of such evidence's *weight*, a concept akin to our "likelihood". Also related is Tversky and Koehler's (1994) support theory, which

postulates that individuals do not attach beliefs to events but to descriptions of events, so that different descriptions of the same event may trigger different assessments. Tversky and Koheler however characterize such non-extensional probability axiomatically, without deriving it from underlying cognitive frictions as we do here.

In our model, diagnostic scenarios quickly pop to the mind of a decision maker, consistent with the idea – supported in cognitive psychology and neurobiology – that background information is a key input in the interpretation of external (e.g., sensory) stimuli.[5] What prevents the local thinker form integrating all other scenarios consistent with the hypothesis, as a Bayesian would do, is assumption A1 of incomplete recall. With complete recall, even our agent is Bayesian. His thinking is System 2 thinking.

The key implication of this setup is that the hypotheses evaluated by the agent themselves influence his assessments by "polluting" his representation of the state space through their effect on the recall and salience of alternative scenarios. This feature is neither shared by existing models of imperfect memory (e.g., Mullainathan 2000, Wilson 2002) nor by models of categorization (e.g., Mullainathan 2002, Mullainathan et al. 2008). In the latter models, there is a first stage in which – irrespective of the hypotheses evaluated by the agent – data provision prompts the choice of a category (akin to a scenario) and a second stage where all hypotheses are evaluated in the same chosen category. In models of categories, the voter observing a candidate not using chopsticks immediately categorizes him as unfamiliar with Asian customs, and within that category he estimates the relative likelihood of qualified and unqualified candidates. In our model,

[5] In the model, background knowledge is summarized by the objective probability distribution $\pi(x)$. This clearly need not be the case. Consistent with memory research, some elements $x \in X$ may get precedence in recall not because they are more frequent but because the agent has experienced them more intensely or because they are easier to recall. Considering these possibilities is an interesting extension of our model.

in contrast, everything happens simultaneously because, on the one hand, the hypotheses themselves affect which scenarios are recalled and, on the other hand, different hypotheses can be represented using different scenarios. In many situations, categorical and local thinking lead to similar assessments of hypotheses, but in many important situations related to KT anomalies, they diverge. Categorical thinking cannot, for example, explain the conjunction and disjunction fallacies, as we discuss below.

4. Biases in Probabilistic Assessments

We measure a local thinker's bias in assessing a generic hypothesis h_1 against an alternative hypothesis h_2 by deriving from expression (9') the odds ratio:

$$\frac{\Pr^L(h_1|d)}{\Pr^L(h_2|d)} = \left[\frac{\sum_{k=1}^{\min(M_1,b)} \Pr(s_1^k|h_1 \cap d)}{\sum_{k=1}^{\min(M_2,b)} \Pr(s_2^k|h_2 \cap d)}\right] \frac{\Pr(h_1|d)}{\Pr(h_2|d)}, \tag{10}$$

where $\Pr(h_1|d)/\Pr(h_2|d)$ is a Bayesian's estimate of the odds of h_1 relative to h_2. One interpretation of (10) is that representations of h_1 and h_2 pop to the agent's mind. The relative likelihood of those representations is captured by the bracketed term. The odds of h_1 are over-estimated if and only if the representation of h_1 is more likely than that of h_2 (the bracketed term is greater than one). Intuitively, a more likely representation induces the agent to over-sample instances of the corresponding hypothesis. Biases arise in our model when a hypothesis is represented with relatively unlikely scenarios.

When b =1, expression (10) becomes:

$$\frac{\Pr^L(h_1|d)}{\Pr^L(h_2|d)} = \left[\frac{\Pr(s_1^1|h_1 \cap d)}{\Pr(s_2^1|h_2 \cap d)}\right] \frac{\Pr(h_1|d)}{\Pr(h_2|d)}, \tag{11}$$

which highlights how *diagnosticity* and *likelihood* of scenarios shape probability estimates. Ceteris paribus, over-estimation of h_1 is the strongest if the diagnostic scenario s_1^1 used to represent h_1 is also the most likely one for h_1, while the diagnostic scenario s_2^1 used to represent h_2 is the least likely one for h_2. In this case, $\Pr(s_1^1 | h_1 \cap d)$ is maximal and $\Pr(s_2^1 | h_2 \cap d)$ is minimal, maximizing the bracketed term in (11). Conversely, under-estimation of h_1 is the strongest if the diagnostic scenario s_1^1 is the least likely one for h_1, while the scenario s_2^1 is the most likely one for h_2.

This illuminates the electoral campaign example of Section 2. Recall that in that example the hypotheses are $h_1 = unqualified$ and $h_2 = qualified$, while their possible scenarios are given by $s \in \{familiar, unfamiliar\}$. Consider the general distribution of candidate types:

Cannot use chopsticks	*Familiar*	*Unfamiliar*
qualified	π_1	π_2
unqualified	π_3	π_4

Table 2.A

We continue to assume that $\pi_1/\pi_3 > \pi_2/\pi_4$, i.e. that being qualified is more likely among familiar than unfamiliar types, so that familiarity with Asian customs is at least slightly informative about qualification. In this case, the diagnostic scenario for $h_1 = unqualified$ is "unfamiliar" while the diagnostic scenario for $h_2 = qualified$ is "familiar". To see this formally, note that $\pi_1/\pi_3 > \pi_2/\pi_4$ implies:

$$\Pr(unqualified | unfamiliar) = \frac{\pi_4}{\pi_2 + \pi_4} > \frac{\pi_3}{\pi_1 + \pi_3} = \Pr(unqualified | familiar),$$

$$\Pr(qualified | familiar) = \frac{\pi_1}{\pi_3 + \pi_1} > \frac{\pi_2}{\pi_4 + \pi_2} = \Pr(qualified | unfamiliar),$$

By A2, these conditions imply that the voter represents h_1 with $(unqualified, unfamiliar)$ and h_2 with $(qualified, familiar)$. In this represented state space, the local thinker estimates $\Pr^L(unqualified) = \pi_4/(\pi_1+\pi_4)$, so that the estimated odds ratio is equal to:

$$\frac{\Pr^L(unqualified)}{\Pr^L(qualified)} = \left[\frac{\pi_4}{\pi_4+\pi_3} \Big/ \frac{\pi_1}{\pi_1+\pi_2}\right]\frac{\pi_3+\pi_4}{\pi_1+\pi_2}, \tag{12}$$

which is the counterpart of (11). The bracketed term is the ratio of the likelihoods of scenarios for low and high qualifications $[\Pr(unfamilar|unqualified)/\Pr(familiar|qualified)]$. The odds that the candidate is unqualified are over-estimated when $\pi_4/\pi_3 > \pi_1/\pi_2$, namely when the share of unfamiliar candidates among the unqualified ones is sufficiently high. In this case, by associating unfamiliarity with low qualifications, the voter forgets that many qualified candidates are also unfamiliar with Asian customs, leading to an over-sampling of unqualified types.

With parameter values in Table 1.A, such over-sampling is strong because π_1 and π_3 are small while π_2 and π_4 are large. This is precisely the case we discussed previously, in which the diagnostic scenario "unfamiliar" used to represent $h_1 = unqualified$ is highly likely [$\pi_4/(\pi_3+\pi_4)$ is large], while the diagnostic scenario "familiar" used to represent $h_2 = qualified$ is unlikely [$\pi_1/(\pi_2+\pi_1)$ is small]. The extreme version of such divergence between diagnosticity and likelihood for $h_2 = qualified$ arises under the following probability distribution of types:

Cannot use chopsticks	*Familiar*	*Unfamiliar*
qualified	$\pi_1 \to 0$	π_2
unqualified	0	π_4

Table 2.B

If $\pi_3 = 0$ and $\pi_1 \to 0$, the diagnosticity of scenarios is preserved because it is still the case that $\pi_1/\pi_3 > \pi_2/\pi_4$. In the limit, the likelihood of the "familiar" scenario for $h_2 = qualified$ becomes zero, so the bias in expression (12) becomes infinite!

In contrast, in the example of Table 1.C, because π_2 and π_3 are small, the most diagnostic and the most likely scenarios coincide for both hypotheses. The extreme version of this case arises when the distribution is:

The Chinese only eat dogs	*familiar*	*Unfamiliar*
qualified	π_1	0
unqualified	0	π_4

Table 2.C

With parameter values in Table 2.C, the bias in expression (12) is zero. When diagnosticity and likelihood coincide, the associations popping up in the agent's mind summarize all relevant cases, entailing no over-sampling and thus no informational loss.

The errors in assessment are particularly high when diagnosticity and likelihood of scenarios are *positively* related for one hypothesis and *negatively* related for the other. When this happens, the representation of the first hypothesis is much more probable than that of the second, leading the agent to over-estimate the probability of the former.

Proposition 1, proved in the Appendix, describes the factors that lead to such asymmetry in the relation between diagnosticity and likelihood across hypotheses.

Proposition 1. Fix h_1, h_2 and denote by S_i the set of scenarios for h_i, $i = 1, 2$. Suppose that $h_2 = \bar{h}_1$ and $S_1 = S_2 = S$. We then have:

a) $s_1^k = s_2^{M-k+1}$ for $k = 1,\ldots,M$ where M denotes the total number of scenarios in S.

b) If $\pi(x)$ is such that $\Pr(s_1^k | h_1 \cap d)$ and $\Pr(s_1^k | h_2 \cap d)$ decrease (increase) in k, then the agent over (under) estimates the odds of h_1 relative to h_2 for every $b < M$. If $b = 1$, one can find a $\pi(x)$ such that such over (under) estimation is arbitrarily large.

c) If $\pi(x)$ is such that $\Pr(s_1^k | h_1 \cap d)$ decreases and $\Pr(s_1^k | h_2 \cap d)$ increases in k, then the maximal factor of under (over) estimation of the odds of h_1 is bounded above by M.

Part a) of Proposition 1 says that competing hypotheses tend to be represented with different scenarios. If h_1 is the negation of h_2, the most diagnostic scenarios for the former are the least diagnostic ones for the latter and vice-versa. Different scenarios necessarily come to mind for the two hypotheses, even if they share the same set of potential scenarios. This result follows from A2 and captures the idea that the agent seeks to build an exemplar representation for each hypothesis, and so must use markedly different scenarios for each. The exemplar of a qualified candidate cannot be the same as that of an unqualified one.

Part b) says that this search for exemplars is a source of pervasive biases if the likelihood ranking of scenarios is the same under both hypotheses. In this case, the use of a highly likely scenario for one hypothesis precludes its use for the competing hypothesis, yielding overestimation of the former. If for the former hypothesis the likelihood and diagnosticity rankings coincide, the bias becomes very strong, potentially infinite. This is the case captured by Table 2.B, where "unfamiliar" is the most likely scenario for both hypotheses but is only used by $h_1 = unqualified$ because it is only diagnostic of that hypothesis. This competition among hypotheses for bringing scenarios to mind is

another crucial way in which hypotheses affect representations in our model. It is a key ingredient in accounting for the biases arising from heuristics.

Part c) instead captures the situation in which the diagnosticity and likelihood of scenarios are positively related for both hypotheses. Biases are now limited (but possibly still large). The largest estimation bias occurs if the likelihood of one hypothesis is fully concentrated on one scenario while the likelihood of the competing hypothesis is fully spread among its M scenarios. In this case, the relative likelihood of the former hypothesis is over-estimated by a factor of M. Hypotheses whose distributions are spread out over a larger number of scenarios are more likely to be underestimated.

4.1 Neglect of Base Rates

Experimental subjects often fail to properly use base rates in assessing probability. KT (1974) gave subjects a personality description of a stereotypical engineer, and told them that he comes from a group of 100 engineers and lawyers, and the share of engineers in the group. Subjects assessed the odds that this person was an engineer or a lawyer. In making this assessment, they mainly focused on the personality description, barely taking the base rates of the engineers in the group into account.

Our model generates base rate neglect. We perform the analysis in a flexible setup based on KT's (1983) famous Linda experiment, to which we return in Section 5 to discuss conjunction fallacies. Subjects are presented with a description of a young woman, called Linda, who is a stereotypical leftist, and in particular was a college activist. They are then asked to check off in order of likelihood the various possibilities of what Linda is today. Subjects estimate that Linda is more likely to be "a bank teller

and a feminist" than merely "a bank teller." We can also use Linda to discuss base rate neglect.

Recall that Linda is described as a former leftist activist (A), and suppose she can be in one of two occupations, bank teller (BT) or social worker (SW) and adhere to one of two current political orientations, feminist (F) or moderate (M). The (unconditional) probability distribution of full descriptions of former activist Linda is displayed below. Crucially, τ and σ are the base probabilities of a bank teller and a social worker in the whole population, respectively.

A (activist)	F (feminist)	M (moderate)
BT (bank teller)	(2/12)τ	(1/12)τ
SW (social worker)	(9/15)σ	(1/15)σ

Table 3.

Table 3 captures two ideas: i) being a former activist reduces the odds of being a bank teller (former activists are only $1/4^{th}$ of all bank tellers but $10/15^{th}s$ of all social workers), and ii) bank tellers are relatively more moderate than social workers (among former activists, moderates are only $1/10^{th}$ of social workers but $1/3^{rd}$ of bank tellers).

A fully local thinker (i.e., $b = 1$) is told that Linda was an activist (i.e., $d = \text{A}$) and asked to assess the probability that she is a bank teller (BT) or a social worker (SW). What comes to his mind? Property ii) of Table 3 implies that the diagnostic scenario for a bank teller is "moderate" (*M*), while that for a social worker is "feminist" (*F*). Formally, $\Pr(BT|A,M) = 5\tau/(5\tau+4\sigma)$, which is greater than $\Pr(BT|A,F) = 5\tau/(5\tau+18\sigma)$. But then, it follows that $\Pr(SW|A,M)$ is smaller than $\Pr(SW|A,F)$. In turn, this implies that a bank teller is represented by (*BT*, *A*, *M*), while a social worker by (*SW*, *A*, *F*), leading to:

$$\frac{\Pr^L(BT|A)}{\Pr^L(SW|A)} = \frac{\Pr(BT,A,M)}{\Pr(SW,A,F)} = \left[\frac{1/3}{9/10}\right]\frac{3}{8}\frac{\tau}{\sigma}. \quad (13)$$

As in (11), the right-most term in (13) is the Bayesian odds ratio, while the bracketed term is the ratio of the two representations' likelihoods. The bracketed term is smaller than one, implying not only that the local thinker under-estimates the odds of Linda being a bank teller, but that he also neglects the information contained in the population odds of a bank teller τ/σ. Even if τ/σ is high, the local thinker under-weights the base rate by a factor of (1/3)/(9/10) = 10/27 relative to the Bayesian assessment.

In our model, neglect of base rates arises because the data d = A skews the agent's recall and thus probability judgement in favour of "social worker", activating in the agent's mind many instances of social workers (the former activists and now feminist), but only a few instances of bank tellers (the former activists and now moderate). This leads to an over-representation of social workers in the agent's mind as he forgets that, among former activists, many bank tellers are feminist.[6] In this sense, our model shows that one effect that KT attribute to agents' use of non-probabilistic logic or heuristics can be rationalized as the result of subjects' limited ability to represent and recall scenarios.

4.2 Insensitivity to Predictability

Various experiments show that people often fail to take into account the reliability of the evidence used in making probabilistic judgements, which are often heavily shaped by scarcely informative data. In one study, KT (1974) presented subjects with descriptions of the performance of a student-teacher during a particular practice lesson.

[6] To allow Table 3 to also illustrate the conjunction fallacy, we assumed that bank teller triggers the least likely scenario of "moderate." Unlike the conjunction fallacy, base rate neglect does not require the difference between diagosticity and likelihood.

Some subjects were asked to evaluate the quality of the lesson, other subjects were asked to predict the standing of each student-teacher five years after the practice lesson. The judgements made under the two conditions were identical, irrespective of subjects' awareness of the limited predictability of teaching competence five years later on the basis of a single trial lesson.

The electoral campaign example of Sections 2 and 3 already showed that local thinkers can over-react to scarcely informative, but diagnostic, evidence. To see this in the context of KT's experiments, suppose that a local thinker assesses the quality of a candidate based on the latter's job talk at a university department. There are three dimensions: the candidate' quality, which can be high (H) or low (L), the quality of his talk, which can be good (GT) or bad (BT), and his expressive ability, which can be articulate (A) or inarticulate (I). The distribution of these characteristics is as follows:

Good Talk (GT)	Inarticulate (I)	Articulate (A)
High Quality (H)	0.005	0.255
Low Quality (L)	0.005	0.235

Table 4.A

Bad Talk (BT)	Inarticulate (I)	Articulate (A)
High Quality (H)	0.235	0.005
Low Quality (L)	0.255	0.005

Table 4.B

In tables 4.A and 4.B, the quality of the talk is highly correlated with expressive ability, but the latter dimension is mildly informative of the candidate's quality. Tables 4.A and 4.B are admittedly extreme, but their similarity to Table 2.B allows us to illustrate the parallel between insensitivity to predictability and the electoral campaign example of Section 3.

Since in Tables 4.A and 4.B the candidate's expressive ability is diagnostic of his quality, after listening to the talk, the local thinker represents low quality candidates as inarticulate, and high quality ones as articulate. The local thinker then assesses:

$$\frac{\Pr^L(H|GT)}{\Pr^L(L|GT)} = \frac{\Pr(H,GT,A)}{\Pr(L,GT,I)} = 51$$
$$\frac{\Pr^L(H|BT)}{\Pr^L(L|BT)} = \frac{\Pr(H,BT,A)}{\Pr(L,BT,I)} = 0.019$$

The local thinker grossly over-estimates the quality of the candidate after a good talk and under-estimates it after a bad talk. Indeed, in our example the quality of the talk conveys very little information about the candidate's quality: a Bayesian would estimate Pr(H|GT)/Pr(L|GT) = 1.08 and Pr(H|BT)/Pr(L|BT) = 0.93 !!

Over-reaction to the quality of the talk is due to the agent's quick association of the candidate's quality and expressive ability, which induces him to miss the fact that the latter attribute is scarcely informative. The general principle here is that there is strong over-reaction when data (quality of the talk) are scarcely informative about the target attribute (quality of the candidate), but very informative about an attribute used by the agent to *represent* different hypotheses (expressive ability).

4.3 The Role of Data-Provision

In the previous example(s), biases result from the agent's reaction to data. How exactly does data provision shape a local thinker's estimate? To answer this question, consider again expression (11) and focus on the bracketed term, measuring the local thinker's bias. If no data is provided, i.e. if $d = X$, this bracketed term is equal to:

$$\frac{\Pr(s_1^1|h_1)}{\Pr(s_2^1|h_2)} = \frac{\Pr(s_1^1 \cap h_1)}{\Pr(s_2^1 \cap h_2)} \bullet \frac{\Pr(h_2)}{\Pr(h_1)}, \qquad (14)$$

where s_i^1 is the diagnostic scenario for h_i when no data is given. In (14), the agent's bias is written as the product of two factors: i) the ratio of the probabilities of representations (the first factor) and ii) the ratio of the probabilities of the hypotheses (the second factor). After data provision (i.e. $d \subset X$), equation (14) becomes:

$$\frac{\Pr(\hat{s}_1^1 | h_1 \cap d)}{\Pr(\hat{s}_2^1 | h_2 \cap d)} = \frac{\Pr(\hat{s}_1^1 \cap h_1 \cap d)}{\Pr(\hat{s}_2^1 \cap h_2 \cap d)} \bullet \frac{\Pr(h_2 \cap d)}{\Pr(h_1 \cap d)}, \tag{15}$$

where $\hat{s}_i^1$ is the diagnostic scenario for h_i when *d* is given. Data reduce the bias if (15) is closer to 1 than (14), they raise the bias otherwise. We cannot say a priori which of these cases we are in, but we can think of the role of data as a combination of the two effects.

First, for a given ratio of the probabilities of representations (the first factor), *d* can boost bias by changing the probabilities of hypotheses (the second factor). Only this effect is at work if the initial scenario s_i^1 is also feasible with data (i.e., $s_i^1 \cap d \neq \phi$ for i = 1,2), since in this case representations do not change. Crucially, if representations do not change, neither does the agent's assessment, even if *d* is objectively informative. This first effect of data, then, captures the *under-reaction* of a local thinker. Through this effect, *d* increases the over-estimation of h_1 if the data are informative about h_2 [i.e. $\Pr(h_1 \cap d)/\Pr(h_2 \cap d) < \Pr(h_1)/\Pr(h_1)$], in which case under-reaction boosts the bias for h_1.

The second effect arises instead when the data "destroy" either or both of the initial scenarios (i.e. $s_i^1 \cap d = \phi$ for some i = 1,2), so that the representation of one or both hypotheses *must* change. Only this effect is at work when *d* is uninformative [i.e. $\Pr(h_1 \cap d)/\Pr(h_2 \cap d) = \Pr(h_1)/\Pr(h_1)$]. This effect captures a local thinker's *over-reaction* and enhances over-estimation of h_1 if the new representation of h_1 triggered by the data is

relatively more likely than that of h_2. In this case, data facilitate the recall of instances supporting h_1 relative to h_2, increasing the over-sampling of the former hypothesis.

This last effect can be seen in the example of Section 4.2, where data, consisting of the job candidate's talk, is almost uninformative about his quality. To see why the agent over-reacts to the talk, consider a local thinker's assessment of the candidate's quality before hearing his talk. Tables 4.A and 4.B imply that without data, a high quality candidate is represented as someone articulate and giving good talks, i.e. with (H,GT,A), a low quality candidate as someone inarticulate and giving bad talks, i.e. with (L,BT,I). The local thinker's unconditional assessment is then given by:

$$\frac{\Pr^L(H)}{\Pr^L(L)} = \frac{\Pr(H, GT, A)}{\Pr(L, BT, I)} = 1, \tag{16}$$

which is equal to the Bayesian odds ratio of Pr(H)/Pr(L) = 1. In this case, limited memory does not create biases.

Why does the talk trigger such over-reaction? Suppose that the talk is bad. While this piece of data is fully consistent with the representation of a bad candidate (who is supposed to give bad talks), it "destroys" the representation of a good candidate, relegating it to the rare exemplar of an articulate candidate who occasionally gives a bad talk. This renders other instances of high quality candidates hard to recall for the agent, giving rise to drastic under-estimation of quality. The reverse occurs after a good talk, which "destroys" only the representation of a low quality candidate, leading to over-estimation of quality. The provision of scarcely informative data causes an over-reaction when such data is consistent with the representation of one of the hypotheses while inconsistent with that of the other, leading to an over-sampling of the former.

We conclude this discussion by illustrating that the same idea can explain the base rate neglect example of Section 4.1. Suppose that distribution of Linda types is:

A / NA	F	M
BT	(2/3)(2τ/8) / (1/5)(6τ/8)	(1/3)(2τ/8) / (4/5)(6τ/8)
SW	(9/10)(2σ/3) / (1/2)(σ/3)	(1/10)(2σ/3) / (1/2)(σ/3)

Table 5.

The numbers above the diagonal capture the distribution of former activist Linda types; while the distribution of non activist types (NA) lie below the diagonal.

Suppose that an agent is asked to assess the probability that Linda is a bank teller or a social worker without being given any data. In this case, the agent represents a bank teller as a "non activist and moderate" and a social worker as an "activist and feminist". It is easy to check that (NA,M) is the diagnostic scenario for bank teller while (A,F) is the diagnostic scenario for social worker. As a consequence:

$$\frac{\Pr^L(BT)}{\Pr^L(SW)} = \frac{\Pr(BT, NA, M)}{\Pr(SW, A, F)} = \frac{(2/3)(6\tau/8)}{(9/10)(2\sigma/3)} = \frac{5}{6}\frac{\tau}{\sigma}, \tag{17}$$

an almost correct unconditional probability assessment, given that the population odds ratio is equal to τ/σ.

A comparison of (17) and (13) shows that the evidence that Linda was an activist mutes the impact of base rates by a factor larger than two. This is so because the data destroys the representation of bank teller, which relies on Linda not being an activist, but not that of a social worker. Such asymmetric impact on the hypotheses' representation implies that $d = A$ reduces the agent's ability to recall instances of bank tellers, inducing an over-sampling of social workers and thus a drastic neglect of bank tellers' base rate.

5. Failures of Extensionality

5.1 Conjunction Fallacy

The conjunction rule states that the probability of a conjoined event C&D cannot exceed the probability of event C or D by itself. KT's (1983) Linda experiment, which we have already described and analyzed for other purposes, dramatically demonstrated the conjunction fallacy. Experimental subjects estimated that Linda the former activist is more likely today to be a feminist bank teller than just a bank teller.

In our model, the conjunction fallacy obtains only under the following necessary condition:

Proposition 2 Fix two hypotheses h_1, h_2. Then, $\Pr^L(h_1 \cap h_2) \geq \Pr^L(h_1)$ only if the scenario s_1 with which the agent represents h_1 is not the most likely one.

The conjunction fallacy arises only if the constituent event h_1 is represented with a diagnostic but unlikely scenario. To see why, denote by $s_{1,2}$ the scenario used to represent the conjunction $h_1 \cap h_2$ and by s_1 the scenario used to represent the constituent event h_1. We study the case with no data, but it is easy to extend the argument to the case in which some data is provided. The conjunction rule is violated when:

$$\Pr(s_{1,2} \cap h_1 \cap h_2) \geq \Pr(s_1 \cap h_1), \tag{18}$$

i.e., when the probability of the *represented* conjunction is higher than the probability of the *represented* constituent event h_1. We can rewrite (18) as:

$$\Pr(s_{1,2} \cap h_2 | h_1) \geq \Pr(s_1 | h_1). \tag{19}$$

The conjunction rule is violated if and only if scenario s_1 is less likely than $s_{1,2} \cap h_2$ for hypothesis h_1. Note, though, that $s_{1,2} \cap h_2$ is itself a scenario for h_1, because

$s_{1,2} \cap h_2 \cap h_1$ identifies an element of X. As a consequence, condition (18) holds only if the diagnostic scenario s_1 used to represent h_1 is not the most likely one, which proves Proposition 2.

Consider how the conjunction rule is violated in the Linda example of Section 4. After hearing Linda described as a former activist (i.e., d = A), the agent – whose probability space is displayed in Table 3 – assesses the probabilities that Linda is a "bank teller" and a "feminist bank teller". As discussed previously, the agent picks the "moderate" scenario for the bank teller. Linda the bank teller is thus represented as "former activist, moderate, bank teller". Linda the "feminist bank teller" leaves instead no gaps to be filled and is represented perfectly, even by a local thinker. Using the values of Table 3, the local thinker estimates:

$$\frac{\Pr^L(BT|A)}{\Pr^L(BT,F|A)} = \frac{\Pr(BT,A,M)}{\Pr(BT,A,F)} = \left[\frac{1/3}{1}\right]\frac{3}{2} = \frac{1}{2} < 1. \qquad (20)$$

The conjunction rule is violated. Intuitively, the diagnostic scenario "moderate" used to represent "bank teller" is very unlikely in light of the fact that Linda is a former activist. The term "bank teller" brings to mind a representation that excludes feminist bank tellers because "feminist" is a characteristic disproportionately associated with social workers, which does not then match the image of an exemplar bank teller.

This discussion highlights the key role played by the data. In this example, the conjunction rule is violated not because "bank teller" is represented with the "moderate" scenario *per se*, but because such a scenario is very unlikely given that Linda is a former activist. This is another instance of the effect of data provision discussed in Section 4.3. If d = A were not provided, then, according to Table 5, the unconditional scenario for

bank teller would be "non activist, moderate" (NA,M), while that for a feminist bank teller would be "activist (A). In this case,

$$\frac{\Pr^{L}(BT)}{\Pr^{L}(BT,F)}=\frac{\Pr(BT,NA,M)}{\Pr(BT,A,F)}=\left[\frac{3/5}{10/19}\right]\frac{60}{19}=\frac{18}{5}>1 \qquad (21)$$

Not only is the conjunction rule not violated, but the odds of "bank teller" are over-estimated. Once more, the reason for the conjunction fallacy is that $d = A$ destroys the likely scenario of "formerly non-activist, moderate," with which "bank teller" is represented.

One explanation of the Linda experiment discussed in KT (1983) holds that the subjects, instead of assessing Pr($BT|A$) and Pr($BT,F|A$), intuitively assess the probabilities of Linda being a former activist under the two hypotheses namely Pr($A|BT$) and Pr($A|F,BT$).[7] This error can yield the conjunction fallacy because being feminist can increase the chance of being Linda. Indeed, in our example in Table 5, Pr($A|BT$) = 1/4 < Pr($A|F,BT$) = 10/19.[8] KT (1983) addressed this possibility in some experiments. In one of them, after being told that the tennis player Bjorn Borg had reached the Wimbledon final, subjects were asked to assess whether it was more likely that in the final Borg would lose the first set or whether he would lose the first set but win the match. Most subjects violated the conjunction rule by stating that the second outcome was more likely than the first. Although our model can explain this evidence, a mechanical assessment of

[7] In personal communication, Xavier Gabaix proposed a "local prime" model complementary to our local thinking model. Such model exploits the above intuition about the conjunction fallacy. Specifically, in the local prime model an agent assessing $h_1, \ldots, h_n$ evaluates $\Pr^{L'}(h_i|d) = \Pr(d|h_i)/[\Pr(h_1|d) + \ldots + \Pr(h_n|d)]$.

[8] On problem with this explanation is that it does not elucidate the thought process by which the subject substitutes the target assessment Pr($h|d$) with the assessment Pr($d|h$). Our model can however help shed light on such thought process. This is seen by writing $\Pr(h_1|d)/\Pr(h_2|d) = [\Pr(d|h_1)/\Pr(d|h_2)]*[\Pr(h_1)/\Pr(h_2)]$. In the latter expression, one way in which subjects may mistakenly estimate the odds of h_1 given d with the odds of d given h is if he mis-estimates the base rates of the hypotheses to be equal, i.e. if $\Pr^{L}(h_1) = \Pr^{L}(h_2)$. Although it is beyond the scope of this paper to identify under what conditions this is indeed the case, in virtue of its ability to account for base rates' neglect our model can allow to address this issue. In particular, it is fairly easy to come up with numerical examples where this is actually the case.

Pr($d|h$) cannot. The reason is that Pr(Borg has reached the final| score in the final) is always equal to one, regardless of the final score.

Most important, the conjunction fallacy explanation based on the substitution of Pr($h|d$) with Pr($d|h$) relies on the provision of data d. This story cannot thus explain the conjunction rule violations that occur in the absence of data provision. To see how our model can account for those, consider another experiment from KT (1983). Subjects are asked to compare the likelihoods of "A massive flood somewhere in North America in which more than 1000 people drown" to that of "An earthquake in California causing a flood in which more than 1000 people drown". Most subjects find the latter event, which a special case of the former, to be nonetheless more likely.

To analyze this experiment, the state space can be described as having three dimensions: the type of flood, which can either be severe (S) or mild (M), the cause of flood, which can either be a earthquake (E) or a tornado (T), and the location of the flood, which can either be California (C) or the rest of North America (NC). The distribution in the state space has the following features:

M / S	E	T
C	$(1-x)e_C$ / xe_C	$t_C/2$ / $t_C/2$
NC	$e_{NC}/2$ / $e_{NC}/2$	$(1-z)t_{NC}$ / zt_{NC}

Table 6

e_L and t_L capture the probabilities of an earthquake and a tornado in location L = C, NC, while $x > 1/2$ and $z > 1/2$ are respectively the share of earthquakes causing severe floods in California and of tornados causing severe floods in the rest of North America. All probabilities must add up to 1. Table 6 captures two key features of a subject´s

beliefs: i) earthquakes are sufficiently milder in the rest of North America than in California that they cause fewer severe floods (only 1/2 of earthquakes cause severe floods in North America, x >1/2 earthquakes cause severe floods in California), and ii) tornados are sufficiently milder in California than in the rest of North America that they cause fewer severe floods (only 1/2 of tornados cause severe floods in California, z > 1/2 tornados cause severe floods in the rest of North America). We make the natural assumption that $z > x$, meaning that tornados are more likely to cause severe floods than earthquakes.

Table 6 implies that a severe flood (*S*) is represented with scenario (*T,NC*), namely as a severe flood caused by a tornado in the rest of North America because $\Pr(S|T,NC) = z$ > $\Pr(S|E,C) = x$ > $\Pr(S|T,C) = \Pr(S|E,NC) = 1/2$. The event "Severe flood caused by an earthquake in California" instead uniquely identifies the scenario (S,C,E). Given these representations, the assessed odds of (S,C,E) relative to (S) are:

$$\frac{\Pr^L(S)}{\Pr^L(S,C,E)} = \frac{\Pr(S,NC,T)}{\Pr(S,C,E)} = \frac{zt_{NC}}{xe_C}. \tag{22}$$

If the probability of disastrous earthquakes in California is sufficiently high relative to that of disastrous tornados in North America, (i.e., $zt_{NC} > xe_C$), the conjunction fallacy arises without data. Intuitively, although tornadoes mainly cause mild floods, they are a prototypical cause of floods. Hence, severe floods are represented as being caused by tornadoes, disregarding that an earthquake in California can cause a very disastrous flood.

5.2 Disjunction Fallacy

According to the disjunction rule, the probability attached to an event *A* should be equal to the total probability of all events whose union is equal to *A*. Fischoff, Slovic and

Lichtenstein (1979) were the first to document the violation of the disjunction rule experimentally. They asked car mechanics, as well as lay people, to estimate the probabilities of different causes of a car's failure to start. They document that on average the probability assigned to the residual hypothesis – "The cause of failure is something other than the battery, fuel system, or the engine" – went up from 0.22 to 0.44 when that hypothesis was broken up into more specific causes (e.g. the starting system, the ignition system). Respondents, including most remarkably experienced car mechanics, discounted hypotheses that were not explicitly mentioned. The under-estimation of implicit disjunctions such as residual hypotheses has been documented in many other experiments and is the key assumption behind Tversky and Koehler's (1994) support theory.

To see whether local thinking can rationalize such disjunction fallacy, compare the assessment of hypothesis h_1 with the assessment of hypotheses $h_{1,1}$ and $h_{1,2}$ where $h_{1,1} \cup h_{1,2} = h_1$. From equation (10), the implicit disjunction h_1 is underestimated when:

$$\Pr(s_{1,1}^1 \cap h_{1,1}) + \Pr(s_{1,2}^1 \cap h_{1,2}) > \Pr(s_1^1 \cap h_1), \tag{23}$$

i.e., when the probability of its representation $s_1^1 \cap h_1$ is smaller than the sum of the probabilities of the representations $s_{1,1}^1 \cap h_{1,1}$ and $s_{1,2}^1 \cap h_{1,2}$ of $h_{1,1}$ and $h_{1,2}$, respectively. A sufficient condition for (24) to hold is that:

$$s_1^1 \cap h_1 \in \left\{ s_{1,1}^1 \cap h_{1,1}, s_{1,2}^1 \cap h_{1,2} \right\}, \tag{24}$$

that is, at least one of hypotheses $h_{1,1}$ and $h_{1,2}$ has the same representation of the implicit disjunction h_1. The appendix shows that this must always be true, which thus implies:

Proposition 3 For any hypotheses h_1, $h_{1,1}$ and $h_{1,2}$ with $h_{1,1} \cup h_{1,2} = h_1$, a local thinker assessment satisfies $\Pr^L(h_{1,1}) + \Pr^L(h_{1,2}) \geq \Pr(h_1)$.

Local thinking leads to underestimation of implicit disjunctions. Intuitively, unpacking a hypothesis h_1 into its constituent events reminds the local thinker of elements of h_1 he would otherwise fail to integrate into his representation. In Proposition 3, the inequality is weak to allow for the possibility that the local thinker's representation is perfect (because for instance h_1 identifies a single element or because b is large). Generically, the inequality is strict when the agent's representations are imperfect.

Consider the following version of the car mechanic experiment. There is only one dimension, the cause of a car's failure to start (i.e., K=1) so that $X \equiv \{battery, fuel, ignition\}$, where *fuel* stands for "fuel system" and *ignition* stands for "ignition system." Assume without loss of generality that $\Pr(battery) > \Pr(fuel) > \Pr(ignition) > 0$. The agent is initially asked to assess the likelihood that the car's failure to start is not due to battery troubles. That is, he is asked to assess the hypotheses $h_1 = \{fuel, ignition\}$, $h_2 = battery$. Since K= 1, there are no scenarios to fit. Yet, since the implicit disjunction $h_1 = \{fuel, ignition\}$ does not pin down an exact value for the car's failure to start, by criterion (8') in Appendix 1 the agent represents it by selecting its most likely element, which is *fuel* by assumption. The local thinker then attaches the probability:

$$\Pr^L(h_1) = \frac{\Pr(fuel)}{\Pr(fuel) + \Pr(battery)} \tag{25}$$

to the cause of the car's failure to start being other than *battery* when this hypothesis is formulated as an implicit disjunction.

Now suppose that the implicit disjunction h_1 is broken up into its constituent elements, $h_{1,1}$ = *fuel* and $h_{1,2}$ = *ignition* (e.g., the individual is asked to separately assess the likelihood that the car's failure to start is due to ignition troubles or to fuel system

troubles). Clearly, the local thinker represents $h_{1,1}$ by *fuel* and $h_{1,2}$ by *ignition*. As before, he represents the other hypothesis h_2 by *battery*. The local thinker now attaches greater probability to the car's failure to start being other than the battery because:

$$\begin{aligned} &\Pr^L(ignition) + \Pr^L(fuel) = \frac{\Pr(ignition) + \Pr(fuel)}{\Pr(ignition) + \Pr(fuel) + \Pr(battery)} \\ &> \Pr^L(h_1) = \frac{\Pr(fuel)}{\Pr(fuel) + \Pr(battery)} \end{aligned} . \tag{26}$$

In other words, we can account for the observed disjunction fallacy.

6. Conclusion

We have presented a simple model of System 1 in which the agent perceives some data, and combines it with information retrieved from memory to evaluate a hypothesis. The central assumption of the model is that, in the first instance, information retrieval from memory is both *limited* and *selected*. Some, but not all, of the missing scenarios come to mind of the decision maker. Moreover, the hypothesis in question primes the selective retrieval of scenarios from memory, with those most predictive of the hypothesis itself being retrieved first. We showed that this simple model accounts for a significant number of experimental results documented by Kahneman and Tversky, most of which are related to the representativeness heuristic. In particular, the model can explain the conjunction and disjunction fallacies exhibited by experimental subjects.

To explain the evidence, we took a narrow view of how recall of various scenarios takes place. In reality, many other factors affect recall. Both availability and anchoring heuristics described by Kahneman and Tversky bear on how scenarios come to mind, but through mechanisms other than those we elaborated.

Perhaps, at a more general level, our model suggests a somewhat different view of heuristics, and of System 1 vs System 2 thinking. From our perspective, intuition and reasoning are not two different modes of thought. Rather, they differ in what is retrieved from memory to make an evaluation. In the case of intuition, the retrieval is not only quick, but also partial and selective. In the case of reasoning of the sort studied by economists, the retrieval is complete.

Indeed, in economic models, we typically think of people receiving limited information from the outside world, but then combining it with everything they know to make evaluations and decisions. The point of our model is that, at least in making quick decisions, people do not bring everything they know to bear on their decisions. Only some information is automatically recalled from passive memory, and – crucially to understanding the world – the things that are recalled might not even be the most useful. Heuristics, then, are not limited decisions. They are decisions like all the others, but based on limited and selected inputs from memory. System 1 and System 2 are examples of the same mode of thought; they differ in what comes to mind.

7. Appendix 1: Generalizing the definition of scenarios

Consider now hypotheses constrain some dimensions of X to be in a certain set without necessarily constraining them to take specific values as in expression (7). Such general hypotheses take the form:

$$\left\{x \in X \middle| x_i \in H_i \quad for \quad some \quad H_i \subset X_i \quad and \quad some \quad i \in I \quad for \quad I \subseteq [1,...,K]\right\} \quad (27)$$

In the above definition, I is the set of dimensions constrained in the hypothesis and H_i is the admissible set specified in the hypothesis for each dimension $i \in I$. Dimensions $i \notin I$ are left completely free in the hypothesis. Note that the hypotheses of expression (7) are special cases of the hypotheses above when H_i is a singleton for every $i \in I$.

To operationalize our definition of scenario in the case of the general hypotheses above, we assume that the agent transform a hypothesis of type (27) into a hypothesis of type (7) by filling specific values in each of the sets H_i for every $i \in I$. At the same time, the agent fills the remaining dimensions (i.e. those left completely free in the hypothesis) by selecting a scenario fulfilling Definition 1. As a result, the agent must now choose not only how to fill a scenario (i.e. the dimensions unrestricted in the hypothesis), but also how to fill the dimensions left unrestricted in the hypothesis. We assume that an agent with $b = 1$ does that by solving the optimization problem:

$$\max_{(x_i \in H_i)_{i \in I}, s \in S} \Pr\left[(x_i)_{i \in I}, (x_i)_{i \in E} \middle| s \cap d\right], \quad (8')$$

where E is a subset of I containing all dimensions that are constrained by equality in the hypothesis. When all the constrained dimensions in the hypothesis are constrained with equality, formally $E = I$, expression (8') boils down into (8). If some dimensions are instead constrained but not by equality, then the agent selects specific values for them in their admissible range so as to maximize the probability that, conditional on the scenario selected and the data, the hypothesis under scrutiny is true. Assumption (8') captures the idea that dimensions explicitly mentioned in the hypothesis are selected to maximize the probability of the latter.[9] It is easy to check that a solution to problem (8') always exists.

[9] We could assume that filling gaps in hypotheses taking the form described in (27) is equivalent to selecting scenarios, that is the agent may maximize (8) subject to selecting scenarios $s \in h_r \cap d$. Although our main results would still hold, in this case all scenarios $s \in h_r \cap d$ would be equally diagnostic, as expression (8) would always be equal to 1. Assumption (8') captures the intuitive idea that the agent also orders the diagnosticity of elements belonging to ranges explicitly mentioned in the hypothesis itself.

With assumption (8') all the results of the paper are generalized to hypotheses taking the form in (27). The only caveat is that in this case the representation $s_r^1 \cap h_r \cap d$ of hypothesis h_r should be read as the intersection of the set identified by the specific values chosen by the agent for representing h_r with the scenario chosen as well as with the data.

8. Appendix 2: Proofs

Proof of Proposition 1. Consider claim a) first. If $h_2 = \bar{h}_1$, the diagnosticity ranking of $s \in S$ for h_1 follows $\Pr(h_1 \cap d|s) = \Pr(h_1 \cap d \cap s)/\left[\Pr(h_1 \cap d \cap s) + \Pr(h_2 \cap d \cap s)\right]$. The diagnosticity ranking of $s \in S$ for h_2 follows $\Pr(h_2 \cap d|s) = 1 - \Pr(h_1 \cap d|s)$. Evidently then, the diagnosticities of scenarios for the two hypotheses are perfectly negatively correlated, formally $s_1^k = s_2^{M-k+1}$ for $k = 1,\ldots,M$.

We now turn to claim b). At any given $b < M$, h_1 is represented with scenarios $\left(s_1^k\right)_{k \le b}$, while h_2 is represented with $\left(s_1^{M+1-k}\right)_{k \le b}$. As such, the odds of h_1 are over-estimated at b if and only if

$$\sum_{k=1}^{b} \Pr(s_1^k | h_1 \cap d) \ge \sum_{k=1}^{b} \Pr(s_1^{M+1-k} | h_2 \cap d) \tag{28}$$

Suppose now that $\Pr(s_1^k | h_1 \cap d)$ and $\Pr(s_1^k | h_2 \cap d)$ strictly decrease in k. Then, one can easily show that the above condition is met for every $b < M$. Suppose in fact that for a certain $b^* < M$ the above condition is not met. That is, suppose that

$$\sum_{k=1}^{b^*} \Pr(s_1^k | h_1 \cap d) < \sum_{k=1}^{b^*} \Pr(s_1^{M+1-k} | h_2 \cap d) \tag{29}$$

Then, at some $b^{**} \le b^*$, it must be the case that $\Pr(s_1^{b^{**}} | h_1 \cap d) < \Pr(s_1^{M+1-b^{**}} | h_2 \cap d)$. But then, since $\Pr(s_1^k | h_1 \cap d)$ and $\Pr(s_1^k | h_2 \cap d)$ strictly decrease in k, it must also be the case that $\Pr(s_1^b | h_1 \cap d) < \Pr(s_1^{M+1-b} | h_2 \cap d)$ for all $b \le b^*$. The same property implies that $\Pr(s_1^b | h_1 \cap d) < \Pr(s_1^{M+1-b} | h_2 \cap d)$ for all $b > b^*$. But then, this implies that (29) must hold for all $b > b^*$, including $b = M$, which is inconsistent with the fact that:

$$\sum_{k=1}^{M} \Pr(s_1^k | h_1 \cap d) = \sum_{k=1}^{M} \Pr(s_1^{M+1-k} | h_2 \cap d) = 1 \tag{30}$$

must necessarily hold. Hence, if $\Pr(s_1^k | h_1 \cap d)$ and $\Pr(s_1^k | h_2 \cap d)$ strictly decrease in k condition (28) must always hold and the odds of h_1 are always (weakly) overestimated. By using the same logic, it is immediate to show that the odds of h_2 are always (weakly) overestimated when $\Pr(s_1^k | h_1 \cap d)$ and $\Pr(s_1^k | h_2 \cap d)$ strictly increase in k. By using the same logic, one can readily show that when $\Pr(s_1^k | h_1 \cap d)$ and $\Pr(s_1^k | h_2 \cap d)$ strictly increase in k. the odds of odds of h_1 are under-estimated for any b.

To see how over-estimation of h_1 may be infinite, consider, in the class of distributions such that $\Pr(s_1^k | h_1 \cap d)$ and $\Pr(s_1^k | h_2 \cap d)$ decreases in k a distribution where $\Pr(s_1^1 \cap h_1 \cap d) = \Pr(h_1 \cap d)\left[1-\varepsilon^2 \frac{1-\varepsilon^{M-1}}{1-\varepsilon}\right]$ and $\Pr(s_1^k \cap h_1 \cap d) = \Pr(h_1 \cap d)\varepsilon^{2(k-2)}$ for all k≥2 and $\Pr(s_1^1 \cap h_2 \cap d) = \Pr(h_2 \cap d)\left[1-\frac{1-\varepsilon^{M-1}}{1-\varepsilon}\right]$ and $\Pr(s_1^k \cap h_2 \cap d) = \Pr(h_2 \cap d)\varepsilon^{(k-2)}$ for all k≥2, where $0 < \varepsilon < 1$. Under this distribution, lower indexed frames are more diagnostic of h_1 because the probability of bundles belonging to h_1 decays much faster with k than that of bundles belonging to h_2. Under both hypotheses, the probability of bundles decreases in k, which implies that this probability distribution belongs to the class where $\Pr(s_1^k | h_1 \cap d)$ and $\Pr(s_1^k | h_2 \cap d)$ decrease in k. Notice that when b =1 the odds of h_1 relative to h_2 are equal to:

$$\frac{\Pr(s_1^1 \cap h_1 \cap d)}{\Pr(s_1^1 \cap h_2 \cap d)} = \frac{\Pr(h_1 \cap d)}{\Pr(h_2 \cap d)} \frac{\left[1-\varepsilon^2 \frac{1-\varepsilon^{M-1}}{1-\varepsilon}\right]}{\varepsilon^{M-2}}$$

For given true odds ratio $\frac{\Pr(h_1 \cap d)}{\Pr(h_2 \cap d)}$, the estimated odds ratio becomes infinite as $\varepsilon \to 0$.

Finally, consider point c). If $\Pr(s_1^k | h_1 \cap d)$ and $\Pr(s_1^{M+1-k} | h_2 \cap d)$ (weakly) decrease in k, the two hypotheses are represented with their most likely frames. Thus, the greatest over estimation of h_1 relative to h_2 is reached when $\Pr(s_1^1 | h_1 \cap d) = 1$ and $\Pr(s_1^M | h_2 \cap d) = 1/M$. That is, when h_1 is concentrated on its representation while the distribution of h_2 is fully dispersed among all frames. In this case, the agent over estimates the odds of h_1 by a factor of M. Accordingly, when $\Pr(s_1^1 | h_1 \cap d) = 1/M$ and $\Pr(s_1^M | h_2 \cap d) = 1$, the agent under estimates the odds of h_1 by a factor of M. To conclude, notice that in those distributions it is indeed the case that k indicates the recall order for h_1 because in both cases the diagnosticity of a frame for h_1 falls in k.

Proof of Proposition 3. The only thing to check to ensure that (25) holds is that s_1^1 is indeed a feasible scenario for at least one of $h_{1,1}$ and $h_{1,2}$. If this is the case, condition (25) intuitively follows by "revealed preference" logic: if s_1^1 is the most diagnostic scenario for h_1, then s_1^1 is going to be the most diagnostic scenario for either $h_{1,1}$ or $h_{1,2}$, because the scenario for h_1 must be the scenario for $h_{1,1}$, for $h_{1,2}$, or for both (as $h_{1,1} \cup h_{1,2} = h_1$). Suppose to the contrary that s_1^1 is a feasible scenario neither for $h_{1,1}$ nor for $h_{1,2}$. By the definition of the scenario, this can only occur if both $h_{1,1}$ and $h_{1,2}$ are such that neither $s_1^1 \cap h_{1,1}$ nor $s_1^1 \cap h_{1,2}$ identifies an element in X. This, however, cannot be the case because by definition s_1^1 is a scenario for h_1, that is $s_1^1 \cap h_1$ identifies one

element in X. By using the distributive property, this implies that $s_1^1 \cap (h_{1'1} \cup h_{1,2}) = s_1^1 \cap h_{1,1} \cup s_1^1 \cap h_{1,2}$ also identifies one element in X. But this implies that s_1^1 must be a scenario for either $h_{1,1}$ or $h_{1,2}$.

References

Fischhoff, Baruch., Paul Slovic, and Sarah Lichtenstein. 1978. "Fault Trees: Sensitivity of Assessed Failure Probabilities to Problem Representation. *Journal of Experimental Psychology: Human Perceptions and Performance*, 4, 330-344.

Griffin, Dale and Amos Tversky. 1992. "The Weighing of Evidence and the Determinants of Confidence." *Cognitive Psychology*, 24, 411-435.

Kahneman, Daniel. 2003. Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review* 93, 1449-1476.

Kahneman, Daniel, and Amos Tversky. 1972. Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430-454.

_____. 1974. "Judgment Under Uncertainty: Heuristics and Biases," *Science*, 185, 1124-1131.

_____. 1983. "Extensional vs. Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment," *Psychological Review*, 91, 293-315.

Maya Bar-Hillel. 1982. "Studies of Representativeness," in D. Kahneman, P. Slovic, and A. Tversky, eds., *Judgement under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press.

Mullainathan, Sendhil. 2000. "Thinking through Categories," mimeo.

_____. 2002. "A Memory-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 117(3), 735-774.

Mullainathan, Sendhil, Joshua Schwartzstein, and Andrei Shleifer. 2008. "Coarse Thinking and Persuasion," *Quarterly Journal of Economics* 123 (2), 577-620.

Popkin, Samuel. 1991. *The Reasoning Voter*. Chicago, IL: University of Chicago Press.

Tversky, Amos, and Derek Koehler. 1994. "Support Theory: A Nonextensional Representation of Subjective Probability," *Psychological Review*, 101, 547-567.

Wilson, Andrea. 2002. "Bounded Memory and Biases in Information Processing," mimeo.