NBER WORKING PAPER SERIES

WHAT COMES TO MIND

Nicola Gennaioli
Andrei Shleifer

Working Paper 15084
http://www.nber.org/papers/w15084

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2009

We are deeply grateful to Josh Schwartzstein for considerable input, and to Pedro Bordalo, Shane Frederick, Xavier Gabaix, Matthew Gentzkow, Daniel Hojman, Elizabeth Kensinger, Daniel Kahneman, Scott Kominers, David Laibson, Sendhil Mullainathan, Giacomo Ponzetto, Drazen Prelec, Mathew Rabin, 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. Shleifer thanks the Kauffman Foundation for research support. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

What Comes to Mind
Nicola Gennaioli and Andrei Shleifer
NBER Working Paper No. 15084
June 2009
JEL No. D03,D81

ABSTRACT

We present a model of judgment 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, intuitive evaluations. Because the automatic retrieval of data from memory is both limited and selected, the agent's evaluations may be severely biased. This framework can account for some of the evidence on heuristics and biases presented by Kahneman and Tversky, including conjunction and disjunction fallacies.

Nicola Gennaioli
CREI
Universitat Pompeu Fabra
Ramon Trias Fargas 25-27
08005 Barcelona (Spain)
ngennaioli@crei.cat

Andrei Shleifer
Department of Economics
Harvard University
Littauer Center M-9
Cambridge, MA 02138
and NBER
ashleifer@harvard.edu

1. Introduction

Since the early 1970s, Daniel Kahneman and Amos Tversky (hereafter KT 1972, 1974, 1983) published a series of remarkable experiments documenting significant deviations from the 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 the experimental evidence.

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 judgment, based on what comes to mind[3].

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 judgment in light of what he is thinking about. Our approach is 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 an object), the decision maker recalls missing scenarios when he evaluates a hypothesis. Kahneman

[2] Partial exceptions include Mullainathan (2000), Griffin and Tversky (1992), and Tversky and Koehler (1994), to which we return in Section 3.3.
[3] Affect and emotion, two commonly noted aspects of system 1, play no role in our analysis.

and Frederick (2005) describe how psychologists think about this process: "The question of why thoughts become accessible – why particular ideas come to mind at particular times – has a long history in psychology and encompasses notions of stimulus salience, associative activation, selective attention, specific training, and priming (p. 271)."

In our model of this process, what is accessible from memory in the first instance – what comes to mind – 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. We model such accessibility by specifying that scenarios come to mind in order of their representativeness, defined as their ability to predict the hypothesis being evaluated relative to other hypotheses[4]. This assumption formalizes some key properties of KT's representativeness heuristic, giving to the latter a specific cognitive content in terms of limited and selective memory. In this model, when the agent 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 an example illustrating our approach and the basic intuition for our results. In Section 3, we present the formal model, and relate 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

[4] We used a different term than representativeness in an earlier draft, since Kahneman and Tversky appear to have a broader idea in mind. Nevertheless, several readers suggested that we use represenativeness, since our definition of represenativeness is close to several observations made by Kahneman and Tversky.

extensionality, namely the conjunction and disjunction fallacies. Section 6 concludes.

2 An Example: Intuitive Reasoning in an Electoral Campaign

Popkin (1991) argues that intuitive reasoning plays a key role in politics, and can help explain the significance that ethnic voters in America attach to the candidates' knowledge of their customs. "In 1972, during New York primaries, Senator George McGovern of South Dakota was courting the Jewish vote, trying to demonstrate his sympathy for Israel. As Richard Reeves wrote for *New York* magazine in August, 'During one of McGovern's first trips into the city he was walked through Queens by city councilman Matthew Troy and one of their first stops was a hot dog stand. "Kosher?" said the guy behind the counter, and the prairie politician looked even blanker than he usually does in big cities. "Kosher!" Troy coached him in a husky whisper. "Kosher and a glass of milk," said McGovern.'" (p. 2).

Popkin argues that such seemingly minor errors may matter a lot in elections because voters judge candidates intuitively. Popkin suggests that in many – perhaps most – cases, such intuitive assessments work pretty well. Yet this leaves open the question: under what circumstances do intuitive assessments lead to biases? To show how our model may shed light on this question, we consider an example in which intuitive reasoning works well, and then return to hotdogs.

Suppose for concreteness that McGovern declared at a Jewish campaign event that Israel was the aggressor in the 1967 war. Suppose that a voter only wants to assess the probability that the candidate is qualified, but thinks most immediately of the candidate's familiarity with his concerns. 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 calling Israel the aggressor, is described in Table 1.A[5]:

Calls Israel aggressor in 1967 war		Familiarity	
		familiar	*unfamiliar*
candidate's qualification	*qualified*	0.15	0.025
	unqualified	0.025	0.8

Table 1.A

Table 1.A captures two ideas: i) calling Israel the aggressor in the 1967 war is very informative about his unfamiliarity with concerns of a Jewish voter (82.5% of the candidates who say this are unfamiliar), and ii) unfamiliarity in turn is very informative about qualification, at least for a Jewish voter (relative to a prior of 1/2 before calling Israel the aggressor). The latter property is reflected in the qualification estimate of a Bayesian voter, which is equal to:

$$\Pr(\textit{qualified}) = \Pr(\textit{qualified}, \textit{familiar}) + \Pr(\textit{qualified}, \textit{unfamiliar}) = 0.175, \qquad (1)$$

where we suppress conditioning on "calling Israel the aggressor". The Bayesian reduces his assessment of qualification sharply because the blunder is so informative.

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

[5] Throughout the analysis we take the basic database of associations (in this example, Table 1.A) as given. One could alternatively specify a very rich and high-dimensional database and endogenously derive a reduced database such as that in Table 1.A, where the agent represents specific hypotheses. One thought process consistent with our model might work as follows. In the first stage, the hypotheses to be tested themselves determine some dimensions of the space. In the current example, hearing about a candidate makes the "qualification" dimension salient to the voter, which pins down the rows of Table 1.A. Then the candidate's statement about Jewish issues brings to mind familiarity with Jewish concerns, namely the columns of Table 1.A. More generally, the agent in our model may fill in scenarios using the dimensions with respect to which the data are quite informative. As we shall see, the fact that the data are informative about the dimension defining the scenarios does not imply that the data are informative about the hypotheses tested by the agent, which is the very source of biases in our model.

qualified and unqualified candidates can be either familiar or unfamiliar with his concerns. 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 voter's mind are examples of *representative*, or stereotypical, qualified and unqualified candidates.

We model this idea by assuming that the voter fits the most *representative* familiarity level – or "scenario" – for each level of qualification of the candidate. We formally define the representative scenario as the familiarity level that best predicts, i.e. is *relatively* more associated with, the respective qualification level. These representative scenarios for a qualified and an unqualified candidate are then given by:

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

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

In Table 1.A, this means that a qualified candidate evokes examples of candidates who are familiar, but an unqualified candidate evokes candidates who are unfamiliar.[6] Qualification and familiarity are associated in the stereotypical qualified candidate. This reduces the voter's actively processed information to the circled diagonal below:

Calls Israel aggressor in 1967 war		Familiarity	
		familiar	*unfamiliar*
qualification of candidate	*qualified*	0.15	0.025
	unqualified	0.025	0.8

Table 1.B

[6] Indeed, Pr(*qual*|*familiar*)=.86>.036=Pr(*qual*|*unfamiliar*). The reverse is true for an unqualified candidate.

The local thinker' assessment uses only stereotypical qualified and unqualified candidates, so his assessment (indicated by superscript L) is equal to:

$$\Pr^{L}(qualified) = \frac{\Pr(qualified, familiar)}{\Pr(qualified, familiar) + \Pr(unqualified, unfamiliar)} \approx 0.158 \qquad (4)$$

Comparing (4) with (1), we see that a local thinker does almost as well as a Bayesian, because stereotypes capture a big chunk of the respective hypotheses' probabilities. In reality as well as in the voter's mind, familiarity and qualification largely go together.

The same idea, however, suggests that in some cases local thinkers make very biased assessments. Return to the example of a candidate's ignorance that drinking milk with hotdogs is not kosher. Suppose that the distribution of candidate types is:

Drinks milk with a hotdog		Familiarity	
		familiar	*unfamiliar*
Qualification of candidate	*qualified*	0.024	0.43
	unqualified	0.026	0.52

Table 1.C

As in the previous case, in Table 1.C the candidate's drinking milk with hotdogs is very informative about his unfamiliarity with the concerns of Jewish voters, but now such unfamiliarity is extremely uninformative about the candidate's qualification (all relative to a prior of 1/2). Indeed, 95% of the candidates do not know the rules of kashrut, including the vast majority of both the qualified and the unqualified ones. In this example, a Bayesian assesses Pr(*qualified*) = 0.454; he realizes that drinking milk with a hotdog is nearly irrelevant for qualification.

The local thinker, in contrast, still associates familiarity with qualification because the stereotypical qualified candidate is still one familiar with his concerns. Formally, the scenario "familiar" yields a higher probability of the candidate being qualified (.024/.05 =

.48) than the scenario "unfamiliar" (.43/.95 = .45). Likewise, the scenario unfamiliar yields a higher probability of the candidate being unqualified (.55) than the scenario familiar (.52). The local thinker then estimates:

$$\Pr^{L}(qualified) = \frac{\Pr(qualified, familiar)}{\Pr(qualified, familiar) + \Pr(unqualified, unfamiliar)} \approx 0.044 \quad (5)$$

which differs from the Bayesian's assessment by a factor of nearly 10! In contrast to the previous case, the local thinker grossly over-reacts to news and mis-estimates probabilities. Now local thinking generates a massive loss of information and bias.

Why this difference in the examples? After all, in both examples the stereotypical qualified candidate is familiar with the voter's concerns, while the stereotypical unqualified candidate is unfamiliar since, in both cases, familiarity and qualification are positively associated in reality. In the initial, more standard, example, the candidate's familiarity is a good proxy for qualification. Formally, almost all qualified candidates are familiar and unqualified ones are unfamiliar, so stereotypical qualified and unqualified candidates are both extremely common. When the stereotypes of qualified and unqualified candidates are not only representative but also likely, the local thinker's bias is kept down. As a consequence, intuitive reasoning delivers good results.

In the second example, in contrast, familiarity is a poor proxy for qualification. Formally, the bulk of both qualified and unqualified candidates are unfamiliar with the voter's concerns, which implies that the stereotypical qualified candidate (familiar with concerns) is very uncommon while the stereotypical unqualified candidate is very common. Hence, although unfamiliarity with the rules of kashrut does not affect much the true probability that the candidate is qualified, it is inconsistent with the voter's stereotypical qualified candidate. But then, by focusing only on the stereotypical

candidates, the local thinker drastically underestimates qualification because he forgets that many qualified candidates are also unfamiliar with the rules of kashrut!

As noted by Popkin (1991), voters fit political facts using candidates' personal data because those data allow them to map the candidate into a stereotypical candidate, which is precisely what happens in our model. In our model, this process leads to good judgments in situations where the dimension defining the stereotype (familiarity) is quite informative about qualification (Table 1.A), while it leads to very biased judgments in situations where the dimension defining the stereotype is scarcely informative about the target assessment of qualification (Table 1.C). We capture this effect by the distinction between the representativeness and likelihood of scenarios. This distinction also plays a key role in rationalizing the biases generated by the use of heuristics.

These considerations establish a further connection of our work to research in psychology, namely the idea of attribute substitution. According to Kahneman and Frederick (2005, p. 269), "When confronted with a difficult question, people may answer an easier one instead and are often unaware of the substitution." Instead of answering a hard question "is the candidate qualified?," the voter answers an easier one, "is he familiar with my concerns?" We show that such attribute substitutions might occur because, rather than thinking about all possibilities, people think in terms of stereotypical candidates, which associate qualification and familiarity. In many situations, such substitution works, as in our initial example where familiarity is a good stand-in for qualification. But in some situations, the answer to a substitute question is not the same as the answer to the first, as when lots of candidates unfamiliar with the rules of kashrut are nonetheless qualified. It is in those situations that intuitive reasoning leads to biased

judgment, as our analysis seeks to show.

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 $X \equiv X_1 \times ... \times X_K$ is a finite state space generated by the product of $K \geq 1$ dimensions and the function $\pi : X \to [0,1]$ maps each element $x \in X$ into a probability $\pi(x) \geq 0$ such that $\sum \pi(x) = 1$. In the example of Section 2, the dimensions of X are the candidate's qualification, his familiarity with voter concerns and his blunder (i.e., K=3), the elements $x \in X$ are candidate types and the probability measure 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, both h_r (r = 1,..., N) and d are subsets of X. If the agent receives no data, then $d = X$: nothing is ruled out. Hypotheses are exhaustive but may be non-exclusive. In (X, π), the probability of h_r given d is determined by Bayes' rule as:

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

In our example, (1) follows from (6) since in Table 1.A the probabilities are normalized by Pr(calls Israel aggressor). 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 elementary events (or 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 agent recalls for each hypothesis the most "representative" elements. 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 d, let M_r denote the number of elements in $h_r \cap d$, $r = 1,\ldots,N$. The agent represents each $h_r \cap d$ using a number $\min(M_r, b)$ of elements $x \in h_r \cap d$, where $b \geq 1$ is the maximum number of elements the agent can recall.

The set $h_r \cap d$ includes all the elements (representations) consistent with hypothesis h_r and with the data d. Two polar cases are of interest: i) the case of $b = 1$, where 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.[7]

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 $h_r \cap d$ identifies a single element in X, even the representation 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

[7] For a given b, we can alternatively refer to b-level local thinking. A1 is one way to capture limited recall. Our substantive results would not change if we alternatively assumed that the agent discounts the probability of certain elements.

entire set $h_r \cap d$ must be collapsed into a single element. To do so, the agent attributes 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 considers one familiarity scenario for each level of qualification.

A more general formal definition of scenarios is as follows: consider the class of problems where h_r and d specify exact values (rather than ranges) for some dimensions of X. In this class of problems, $h_r \cap d$ takes the form:

$$h_r \cap d \equiv \left\{x \in X \middle| x_i = \hat{x}_i \quad where \quad \hat{x}_i \quad is \quad fixed \quad for \quad some \quad i \in [1,...,K]\right\}, \qquad (7)$$

where $\hat{x}_i$ is the exact value taken by the i-th dimension in the hypothesis or data. The remaining dimensions are unrestricted. This is consistent with the example in Section 2, where hypotheses specify qualification levels, data specify candidate statements, and the remaining familiarity dimension is left completely free. The possible scenarios for the class of hypotheses $h_r \cap d$ in (7) are defined as follows:

Definition 1. Suppose that $h_r \cap d$ fixes the values of $N_r < K$ dimensions in X. Denote by F_r the set of the remaining $K - N_r \geq 0$ free dimensions. If F_r is non empty, a scenario s for $h_r \cap d$ is any event $s \equiv \left\{x \in X \middle| x_t = x_t' \quad for \quad all \quad t \in F_r\right\}$. If F_r is empty, the only scenario for $h_r \cap d$ is $s = X$. S_r is the set of all possible scenarios for hypothesis h_r.

A scenario fills the details missing from the hypothesis and data by identifying a single element in $h_r \cap d$, which we denote by $s \cap h_r \cap d \in X$. How do scenarios come to mind? We assume that the agent represents hypotheses belonging to class (7) as follows:

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

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

where $\overline{h_r}$ is the complement $X \setminus h_r$ in X of hypothesis h_r. When $b > 1$, the agent represents h_r with b most "representative" scenarios s_r^k, $k = 1,\ldots,b$, where scenarios with a lower index k yield a higher value of (8) and where we define $s_r^k = \phi$ for $k > M_r$.

The local thinker represents h_r by recalling only the b most "representative" scenarios, those that are more associated with h_r relative to its complement $\overline{h_r}$. Scenario s_r^1 is the most representative model for the hypothesis h_r in the sense that, together with the data, it maximizes the likelihood $\Pr(h_r | s \cap d)$ of the hypothesis. It is useful to view the intersection of the data, the hypothesis, and the first scenario that comes to mind as a stereotype that the local thinker imagines.

The above formal definition of representativeness has two key properties. First, an attribute is more representative of a hypothesis not only if it is more associated with it, but also if it is less associated with all other hypotheses. That is, an attribute is perceived as more stereotypical if it maximizes the contrast between a hypothesis and its complement. Second, what is stereotypical for one hypothesis is independent of other hypotheses being explicitly evaluated by the agent, as our definition of representativeness only refers to the relationship between a hypothesis and its complement. Still, one key property of A2 is that the way a hypothesis is formulated affects its representation.

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 $h_1,...,h_N$ evaluated by the agent yields:

Definition 2 Given data *d* and hypotheses h_r, $r = 1,\ldots,N$, the agent's representation of any hypothesis h_r is defined as $\widetilde{h}_r(d) \equiv \bigcup_{k=1,...,b} s_r^k \cap h_r \cap d$, and the agent's represented state space $\widetilde{X}$ is defined as $\widetilde{X} \equiv \bigcup_{r=1,...,N} \widetilde{h}_r(d)$.

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 belonging to the class in (7), but it is easy to extend it to general hypotheses which, rather than attributing exact values, restrict the range of some dimensions of *X*. Appendix 1 shows how to do this and to apply our model to the evaluation of these hypotheses as well. The only result in what follows that relies on restricting the analysis to the class of hypotheses in (7) is Proposition 1. As we show in the appendix, all other results, including the role of diagnosticity vs. likelihood as well as the results on the disjunction and conjunction fallacy of Section 5 can be easily extended to fully general classes of hypotheses.

3.2 Probabilistic Assessments by a Local Thinker

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

$$\Pr^L(h_t|d) = \frac{\Pr(\widetilde{h}_t(d))}{\Pr(\widetilde{X})}, \tag{9}$$

which is the probability of the *representation* of h_t divided by that of the represented state space $\widetilde{X}$. One crucial property of (9) is that the assessed probability of a hypothesis depends on the other hypotheses examined in conjunction with it as the latter affects the represented state space and thus the denominator of (9). 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.

Expression (9) highlights the role of local thinking. If $b \geq M_r$ for all $r = 1,..,N$, then $\widetilde{X} = X \cap d$, $\widetilde{h}_t(d) \equiv h_t \cap d$ and (9) boils down to $\Pr(h_t \cap d)/\Pr(d)$, which is the Bayesian's estimate of $\Pr(h_t|d)$. Biases in judgment can only arise when the agent's representations are limited, that is, when $b < M_r$ for some r.

When the hypotheses are exclusive [i.e. $h_t \cap h_r = \phi \ \forall t, r$], (9) can be written as:

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

where $\Pr(s|h_r \cap d)$ is the *likelihood* of scenario s for h_r, or the probability of s when h_r is true. The bracketed terms in (9') 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 (all 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 representativeness.

3.3 Discussion of the Setup and the 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 judgment 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 representativeness 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 use of *representative* but unlikely scenarios drives many of the KT anomalies.

Our approach is also 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 "representativeness", 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, representative 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.[8] 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.[9]

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 analogical thinking (Jehiel 2005) or 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 Jewish voter observing a candidate drinking milk with a hotdog immediately categorizes him as unfamiliar with his concerns, and within that category he estimates the relative likelihood

[8] 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.

[9] Our approach shares some similarities with models of sampling. Stewart et al. (2006) study how agents form preferences over choices by sampling their past experiences, Osborne and Rubinstein (1998) study the determination of equilibrium in strategic settings where players sample the performance of different actions. The focus of these works is very different from judgment under uncertainty. From a general behavioural standpoint, however, the fundamental innovation of our work is to consider the case where agents do not sample scenarios at random but based on their representativeness (or more broadly accessibility) leading them to over-sample certain specific memories and under-sample others.

of qualified and unqualified candidates. He would make a mistake in assessing qualification, but only a small one when virtually all candidates are unfamiliar.

In 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, competing hypotheses are represented using different scenarios. In many situations, categorical and local thinking lead to similar assessments of hypotheses, but in situations closely 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:

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

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 one 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 *representativeness* and *likelihood* of scenarios shape probability estimates. Ceteris paribus, over-estimation of h_1 is the strongest if the representative scenario s_1^1 used to represent h_1 is also the most likely one for h_1, while the representative 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 representative 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 analysis illuminates the electoral campaign example of Section 2. Consider the general distribution of candidate types when *d* = "drinks milk with a hotdog":

Drinks milk with a hotdog	*familiar*	*unfamiliar*
qualified	π_1	π_2
unqualified	π_3	π_4

Table 2.A

We 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 Jewish concerns is at least slightly informative about qualification. This implies that:

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

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

The representative scenario for $h_1 = unqualified$ is then $s_1^1 = unfamiliar$, while the representative scenario for $h_2 = qualified$ is $s_2^1 = familiar$. By A2, 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:

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

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 qualification in the stereotypical candidates, the voter forgets that many qualified candidates are also unfamiliar with the rules of kashrut, leading to an over-sampling of unqualified types.

In the example of Table 1.A, judgments are good because π_2 and π_3 are small, so that the most representative and the most likely scenarios coincide for both hypotheses. The extreme version of this case arises when the distribution is given by Table 2.B:

Calls Israel aggressor in 1967 war	*familiar*	*unfamiliar*
qualified	π_1	0
unqualified	0	π_4

Table 2.B

With parameter values in Table 2.B, the bias in expression (12) is zero. Each hypothesis is fully represented by its stereotype and so local thinking entails no informational loss, leading to a perfect assessment. More generally, even if some probability mass is placed on non-stereotypical candidates, assessments are perfect (or moderately biased) provided that representativeness and likelihood coincide for both hypotheses.

Matters are radically different with the parameter values in Table 1.C, in which case π_1 and π_3 are small while π_2 and π_4 are large. This is precisely the case in which the representative scenario "unfamiliar" used to represent $h_1 = unqualified$ is highly likely [$\pi_4/(\pi_3+\pi_4)$ is large], while the representative scenario "familiar" used to represent $h_2 = qualified$ is unlikely [$\pi_1/(\pi_2+\pi_1)$ is small]. As shown by (12), in this case biases are large. The extreme version of such divergence between representativeness and likelihood for $h_2 = qualified$ arises under the following probability distribution of types:

Drinks milk with a hotdog	*familiar*	*unfamiliar*
qualified	$\pi_1 \to 0$	π_2
unqualified	0	π_4

Table 2.C

If $\pi_3 = 0$, the representativeness of scenarios is preserved because it is still the case that $\pi_1/\pi_3 > \pi_2/\pi_4$. However, as $\pi_1 \to 0$, the likelihood of the "familiar" scenario for $h_2 = qualified$ becomes zero, so the over-estimation factor in expression (12) becomes infinite! In this case, the local thinker's focus on stereotypes induces him to grossly over-sample unqualified candidates, and leads to severe underestimation of qualification.

To summarize, the errors in assessment are particularly high when representativeness 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.

To see in a more general setting how biases arise in our model and what determines their strength, consider the following proposition, which is proved in the Appendix and is restricted to the class of hypotheses described in (7):

Proposition 1. Suppose that the agent evaluates hypotheses h_1, h_2 where $h_2 = \bar{h}_1$ and the set of feasible scenarios for them is the same, namely $S_1 = S_2 = S$. We then have:

1) Representation: scenarios rank in opposite order of representativeness for the two hypotheses, formally $s_1^k = s_2^{M-k+1}$ for $k = 1,\ldots,M$ where M is the number of scenarios in S.

2) Assessment bias:

i) If $\pi(x)$ is such that $\Pr(s_1^k | h_1 \cap d)$ and $\Pr(s_1^k | h_2 \cap d)$ strictly decrease in k (at least for some k), the representativeness and likelihood of scenarios are positively related for h_1, and negatively related for h_2. The agent thus over-estimates the odds of h_1 relative to h_2 for every $b < M$. One can find a $\pi(x)$ so that such over estimation is arbitrarily large. The opposite is true if $\Pr(s_1^k | h_1 \cap d)$ and $\Pr(s_1^k | h_2 \cap d)$ strictly increase in k.

ii) 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, the representativeness and likelihood of scenarios are positively related for both hypotheses. The agent over- or underestimates the odds of h_1 relative to h_2 at most by a factor of M/b.

Proposition 1 usefully breaks down the roles of assumption A.2 and of the probability distribution $\pi(x)$ in generating biases.[10] With respect to representations, A.2 implies that when $h_2 = \bar{h}_1$, the most representative scenarios for h_2 are the least representative ones for h_1 and vice-versa. The difference in the representativeness of the same scenario for the two hypotheses would not be so stark if $h_2 \neq \bar{h}_1$, since in that case h_1 and h_2 could be represented with the same scenario. We stress the case in which

[10] The proof of the Proposition provides detailed conditions on classes of problems in which it is indeed the case that $S_1 = S_2 = S$.

$h_2 = \bar{h}_1$ because, by A2, recall maximizes the contrast in the representation of different hypotheses. Intuitively, the stereotypes for competing hypotheses are different: the stereotype of a qualified candidate is very different from that of an unqualified one.

What does this property of representations imply for biases? Part 2.i) says that this reliance on different stereotypes causes pervasive biases if the likelihood ranking of scenarios is the same under each hypothesis. In this case, the use of a highly likely scenario for one hypothesis precludes its use for the competing hypothesis, leading to overestimation of the former. The resulting bias can even be infinite. This is what happens in Table 2.C, where "unfamiliar" is the most likely scenario for both hypotheses but is only used for $h_1 = unqualified$ because it is only representative of that hypothesis. In general, case 2.i) formally captures situations that are informative about some specific scenarios (the most likely ones), but where those scenarios are in turn not very informative about h_2, h_1 because they are likely under both hypotheses.

In this case, the use of stereotypes or, more generally, of limited representation leads to strong biases. As we shall see, this effect plays a crucial role in accounting for the biases arising from heuristics, especially the disjunction fallacy.

Finally, part 2.ii) captures the case where the representativeness and likelihood of scenarios are positively related for both hypotheses. Biases are now limited (but possibly still quite large) and 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. This implies that hypotheses whose distribution is spread out over a larger number of scenarios are more likely to be underestimated. The

maximum bias falls in b because recall of more scenarios attenuates the impact on assessments of the differences in how hypotheses are spread out across scenarios.

We next show how our model helps rationalize two specific biases related to the representativeness heuristic: base-rate neglect and insensitivity to predictability. We focus on numerical examples, but we also discuss the general ideas behind our results.

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.[11]

Our model generates base-rate neglect as the consequence of agents' assessing probabilities by retrieving stereotypes. We now show how this works 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

[11] The illustrations here are Linda, lawyer-engineer and other examples of base-rate neglect related to the representativeness heuristic (more broadly to association of hypotheses with stereotypes). Base-rate neglect is more general, since it also occurs in judgments that cannot be readily interpreted in terms of representativeness, such as the Cascells et al. (1978) experiment on physicians' interpretation of clinical tests or the KT's blue vs. green cab experiment (KT 1982, p. 154). These latter instances of base-rate neglect are unlikely to be due to the use of stereotypes, although they might well be a product of local thinking. KT (1982, p. 154) describe the differences between these two forms of base-rates neglect.

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."

To discuss base-rate neglect in this context, suppose 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 in Table 3, where 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 of all bank tellers but 2/3 of all social workers), and ii) among former activists, bank tellers are relatively more moderate than social workers (moderates are only 1/10 of social workers but 1/3 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). The Bayesian odds are $(1/4\ \tau)/(2/3\sigma) = (3/8)(\tau/\sigma)$. But what comes to the local thinker's mind? Property ii) of Table 3 implies that the representative scenario for ''bank teller'' is "moderate" (M), that for ''social worker'' is "feminist" (F). Formally, $\Pr(BT|A,M) = 5\tau/(5\tau+4\sigma) > \Pr(BT|A,F) = 5\tau/(5\tau+18\sigma)$ and thus $\Pr(SW|A,M) > \Pr(SW|A,F)$. As a result, ''bank teller'' is represented by (BT, A, M), ''social worker'' by (SW, A, F), leading to:

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

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 some of the information contained in the population odds of a bank teller, τ/σ. The local thinker under-weights the base-rate by a factor of (1/3)/(9/10) = 10/27 relative to the Bayesian assessment.

Neglect of base-rates arises here because the data d = A is more consistent with "feminist" than with "moderate", thereby skewing the agent's recall and thus probability judgment in favor of "social worker". That is, d = A induces retrieval of many instances of formerly activist and feminist social workers, but only a few instances of formerly activist, but now moderate, bank tellers, leading to an over-sampling and thus over-estimation of social workers. In our specific numerical example, bank teller, in contrast to social worker, prompts the recall of a representative but unlikely scenario. As we shall see, this feature is essential for explaining the conjunction fallacy. In the present context, however, although the use of an unlikely scenario renders biases more severe, it is not necessary for base-rate neglect, which is a natural consequence on the agent's use of limited, stereotypical information. 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 reveal people's failure to account for the reliability of the evidence used in making probabilistic judgments, 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. 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 judgments 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 representative, 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	0.25
Low Quality (L)	0.005	0.24

Table 4.A

Bad Talk (BT)	Inarticulate (I)	Articulate (A)
High Quality (H)	0.24	0.005
Low Quality (L)	0.25	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.C shows the parallel between insensitivity to predictability and the electoral campaign example of Section 2.

Since in Tables 4.A and 4.B the candidate's expressive ability is representative 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:

$$\frac{\Pr^L(H|GT)}{\Pr^L(L|GT)} = \frac{\Pr(H,GT,A)}{\Pr(L,GT,I)} = 50$$

$$\frac{\Pr^L(H|BT)}{\Pr^L(L|BT)} = \frac{\Pr(H,BT,A)}{\Pr(L,BT,I)} = 0.02$$

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) = 1.04 and Pr(H|BT)/Pr(L|BT) = 0.96 !!

Once more, over-reaction to the quality of the talk is due to the agent's quick association of the candidate's quality and expressive ability. It is useful to interpret this result in terms of the use of stereotypes. Over-reaction here is due to the fact that the data (quality of the talk) are scarcely informative about the target attribute (quality of the candidate), but very informative about an attribute defining the stereotype for different hypotheses (expressive ability). As in the Linda example, Tables 4.A and 4.B exploit the divergence between representativeness and likelihood to illustrate this phenomenon in the starkest manner, but over-reaction to data is a natural and general consequence of the use of stereotypes, as shown in the section below.

4.3 The Role of Data-Provision

Local thinkers' biases described in Proposition 1 do not rely in any fundamental way on data provision (i.e., they also arise for $d = X$). However, since in the experiments

considered so far biases resulted from the explicit exposure of agents to some piece of data, this is an appropriate place to look more closely at the role of data provision in our model. To do so, consider again (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:

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

where s_i^1 is the representative 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 representative scenario for h_i when d is given. Data lower 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 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 representations do not change (a sufficient condition for representations not to change is that s_i^1 is also feasible with data, i.e., $s_i^1 \cap d \neq \phi$, but $s_i^1 \cap \overline{d} = \phi$ for i=1,2). Indeed, 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* by a local thinker because in this case data provision leaves unchanged the assessment by a local thinker but not that by a Bayesian. Specifically, d

increases the over-estimation of h_1 if and only 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. if $s_i^1 \cap d = \phi$ 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. $\Pr(h_1 \cap d)/\Pr(h_2 \cap d) = \Pr(h_1)/\Pr(h_2)$]. 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 second effect can be seen in the base-rate neglect phenomenon of Section 4.1. Suppose that before receiving any information, the distribution of Linda types is:

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

Table 5.

The entries above the diagonal capture the distribution of former activist Linda types (same as in Table 3); those below the diagonal show the distribution of former non activist types (NA). Obviously, $\tau + \sigma = 1$. The data d = A is informative about the probability that Linda is a bank teller [whose odds fall from τ/σ to $(3/8)*(\tau/\sigma)$], but – as shown in (13) – the local thinker grossly mis-estimates these odds.

To see how this is due to the effect of data on stereotypes, suppose that the agent is asked to assess the probability that Linda is a bank teller or a social worker without

being given any data. It is easy to show that now the agent represents a bank teller as a "non activist and moderate", a social worker as an "activist and feminist". As a result:

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

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

More important, a comparison of (17) to (13) shows that the agent grossly over-reacts to the evidence that Linda was an activist. After seeing *d*=A, the local thinker reduces the odds of bank teller much more than a Bayesian [specifically, by a factor of about (1/3)*(3/8) versus a factor of 3/8]. As discussed above, this over-reaction is due to the fact that the data *d* = A "destroys" the stereotype of bank teller, which relies on Linda being a non-activist, but not that of a social worker, which is instead perfectly consistent with her being a former activist. Such asymmetric impact on the hypotheses' representations imply that the data *d* =A relegate the hypothesis of bank teller to the rare exemplar of a former activist and now moderate Linda. This reduces the agent's ability to recall instances of bank tellers, inducing an over-sampling of social workers and thus a drastic over-reaction to data.

The same effect is at play in the insensitivity to predictability experiment, where a good talk destroys only the stereotype of a bad candidate while a bad talk only that of a good candidate. Over-reaction to the data is particularly severe here because the quality of the talk is scarcely informative about the quality of the candidate.

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.

For simplicity, we only study the conjunction fallacy when b=1 and when the agent is provided no data, but it is possible to relax these assumptions because the fundamental logic of the conjunction fallacy does not rely on them. We now formally consider the class of problems in (7), but in the appendix we prove that Proposition 2 below holds also for general classes of hypotheses.

We focus on the so-called "direct tests", namely when the agent is asked to simultaneously assess the probability of a conjoined event $h_1 \cap h_2$ and of one of its constituent events such as h_1. Denote by $s^1_{1,2}$ the scenario used to represent the conjunction $h_1 \cap h_2$ and by s^1_1 the scenario used to represent the constituent event h_1. In this case, the conjunction fallacy obtains in our model if and only if:

$$\Pr(s^1_{1,2} \cap h_1 \cap h_2) \geq \Pr(s^1_1 \cap h_1), \tag{18}$$

i.e., when the probability of the *represented* conjunction is higher than the probability of the *represented* constituent event h_1. Expression (18) is a direct consequence of (9), as in this direct test the denominators are identical and cancel out. The conjunction fallacy [expression (18)] then arises only under the following necessary condition:

Proposition 2. When b=1, in a direct test of hypotheses h_1 and $h_1 \cap h_2$, $\Pr^L(h_1 \cap h_2) \geq \Pr^L(h_1)$ only if scenario s_1^1 is not the most likely for h_1.

The conjunction fallacy arises only if the constituent event h_1 is represented with a representative but unlikely scenario. To see why, rewrite (18) as:

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

The conjunction rule is violated when scenario s_1^1 is less likely than $s_{1,2}^1 \cap h_2$ for hypothesis h_1. Note, though, that $s_{1,2}^1 \cap h_2$ is itself a scenario for h_1 since $s_{1,2}^1 \cap h_2 \cap h_1$ identifies an element of X. As a consequence, condition (18) only holds if the representative scenario s_1^1 is not the most likely scenario for h_1, which proves Proposition 2. It is then obvious that, whenever a hypothesis h_1 is not represented with the most likely scenario, to induce the conjunction fallacy it is *sufficient* for the agent to test hypothesis h_1 against the conjoined hypothesis $h_1^* = s_1^* \cap h_1$, where s_1^* is the most likely scenario for hypothesis h_1 and $h_1^* \subset h_1$ is the element obtained by fitting such most likely scenario in hypothesis h_1 itself. By construction, in this case $\Pr^L(h_1^*) \geq \Pr^L(h_1)$, so that the conjunction rule is clearly violated.

Consider now one specific instance of the conjunction fallacy in the Linda example from 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 represents Linda the bank teller by picking the "moderate" scenario, whereas Linda

the "feminist bank teller" leaves no gaps to be filled and is represented perfectly, even by a local thinker. Using the values of Table 3, the local thinker estimates:

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

The conjunction rule is violated.

In line with Proposition 2, representativeness and likelihood diverge because the "moderate" scenario used to represent Linda the bank teller is less likely than the "feminist" scenario, as "moderate" is very unlikely in light of the fact that Linda is a former activist. Why does the agent fail to realize this fact? Our answer is that 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 role played by the data. The conjunction rule is violated here 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. 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,

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

Not only is the conjunction rule not violated, but the odds of "bank teller" are over-estimated. This is another instance of the effect of data provision discussed in Section 4.3: the agent violates the conjunction rule because $d = A$ 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($A|F,BT$).[12] 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. However, KT (1983) addressed this possibility in some experiments. In one of them, subjects were provided with the data that the tennis player Bjorn Borg had reached the Wimbledon final, and then 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. Our model can explain this evidence, but a mechanical assessment of Pr($d|h$) cannot. The reason is that Pr(Borg has reached the final| Borg's 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 is a special case of the former, to be nonetheless more likely.

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

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 features of a subject's beliefs: i) earthquakes are sufficiently milder in the rest of North America than in California in 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 $z > x$, so 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 stereotypical cause of floods. Hence, severe floods are represented as being caused by tornadoes, even though agents hold the belief that earthquakes in California can be so severe as to cause much more disastrous floods. The problem, though, is that agents retrieve this belief only if earthquakes and California are explicitly mentioned.

The general idea behind these types of conjunction fallacy is that either the data (Linda is a former activist) or the question itself (floods in North America) bring to mind a representative but unlikely scenario. Besides the examples specifically discussed, this general principle can help explain other types of conjunction rule violations. For example, Kahneman and Frederick (2005) report that subjects estimate the annual number of murders in the state of Michigan to be lower than that in the city of Detroit, which is in Michigan. Our model suggests that this might be explained by the fact that the stereotypical location in Michigan is rural and non-violent, so subjects forget that the more violent city of Detroit is in the state of Michigan as well.

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*. Fischhoff, Slovic

and Lichtenstein (1978) 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 $h_{1,1} \cup h_{1,2} = h_1$ (and obviously $h_{1,1} \cap h_{1,2} = \phi$) by an agent with b=1. It is straightforward to extend the result to the case where b>1. Formally, we compare $\Pr^L(h_1)$ when h_1 is tested against $\overline{h}_1$ with the sum $\Pr^L(h_{11}) + \Pr^L(h_{1,2})$ obtained when $h_{1,1}$ and $h_{1,2}$ are tested against their complement $\overline{h}_1$. The local thinker then violates the disjunction rule in the direction of the Fischhoff et al. experiment provided $\Pr^L(h_{11}) + \Pr^L(h_{1,2}) > \Pr^L(h_1)$.

Define s_1^1, $s_{1,1}^1$, $s_{1,2}^1$ and $s_{\overline{1}}^1$ to be the representative scenarios for hypotheses, h_1, $h_{1,1}$, $h_{1,2}$, and $\overline{h}_1$ respectively. From equation (9), it then follows that the implicit disjunction h_1 is underestimated when:

$$\frac{\Pr(s_{1,1}^1 \cap h_{1,1}) + \Pr(s_{1,2}^1 \cap h_{1,2})}{\Pr(s_{1,1}^1 \cap h_{1,1}) + \Pr(s_{1,2}^1 \cap h_{1,2}) + \Pr(s_{\overline{1}}^1 \cap \overline{h}_1)} > \frac{\Pr(s_1^1 \cap h_1)}{\Pr(s_1^1 \cap h_1) + \Pr(s_{\overline{1}}^1 \cap \overline{h}_1)}. \qquad (23)$$

Equation (23) immediately boils down to:

$$\Pr(s^1_{1,1} \cap h_{1,1}) + \Pr(s^1_{1,2} \cap h_{1,2}) > \Pr(s^1_1 \cap h_1), \tag{23'}$$

meaning that the probability of the representation $s^1_1 \cap h_1$ of h_1 is smaller than the sum of the probabilities of the representations $s^1_{1,1} \cap h_{1,1}$ and $s^1_{1,2} \cap h_{1,2}$ of $h_{1,1}$ and $h_{1,2}$, respectively. The appendix proves that this occurs if the following condition holds:

Proposition 3. Suppose that $b = 1$. In one test, hypothesis h_1 is tested against a set of alternatives. In another test, hypotheses $h_{1,1}$ and $h_{1,2}$, are jointly tested against the same set of alternatives as h_1. Then, if s^1_1 is also a feasible scenario for $h_{1,1}$ and $h_{1,2}$, it follows that $\Pr^L(h_{1,1}) + \Pr^L(h_{1,2}) > \Pr(h_1)$.

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 which he would otherwise fail to integrate into his representation. The sufficient condition for this to occur (i.e., that s^1_1 must be a feasible scenario in the explicit disjunction) is very weak. For example, it is always fulfilled when the representation of the implicit disjunction $s^1_1 \cap h_1$ is contained in a residual category of the explicit disjunction such as "other". The logic for this result is that the condition in Proposition 3 implies that:

$$s^1_1 \cap h_1 \in \left\{ s^1_{1,1} \cap h_{1,1}, s^1_{1,2} \cap h_{1,2} \right\}, \tag{24}$$

that is, at least one of hypotheses $h_{1,1}$ and $h_{1,2}$ has the same representation as the implicit disjunction h_1.

Proposition 3 is directly proved for general hypotheses, not only for those belonging to the class in (7). This allows us to immediately consider the following model

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 $\Pr(battery) > \Pr(fuel) > \Pr(ignition) > 0$. This case meets the conditions of Proposition 3 because now no dimension is left free, so all hypotheses share the same scenario $s = X$.

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*. 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. The logic is the same as that of Proposition 3: the representation of the explicit disjunction adds to the representation of the implicit disjunction (i.e. x = *fuel*) an additional element (i.e. x = *ignition*), which boosts the assessed probability of the explicit disjunction.

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 *selective*. 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 relative to the other hypotheses – the representative scenarios -- being retrieved first. In many situations, such intuitive judgment works well, and does not lead to large biases in probability assessments. But in situations where the representativeness and likelihood of scenarios diverge, intuitive judgment becomes faulty. 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.

Appendix 1: Generalizing the definition of scenarios

Consider general combinations of hypotheses and data constraining some dimensions of X to be in a certain set without necessarily fixing specific values for them as in (7). Now:

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

where I is the set of dimensions constrained by the hypothesis, H_i is the set specified in the hypothesis (and the data) for each $i \in I$. Dimensions $i \notin I$ are left completely free. The class of hypotheses in (7) is a special case of that in (27) when sets H_i are singletons.

To operationalize our definition of a scenario, we assume that the agent transforms a hypothesis of type (27) into a hypothesis of type (7) by filling specific values in each set H_i for every $i \in I$. The agent also fills the dimensions left completely free (i.e. values for $i \notin I$) by selecting a scenario fulfilling Definition 1. We assume that an agent with $b = 1$ does that by solving:

$$\max_{x_I, s} \Pr\left[x_I \middle| s \cap d\right], \quad (8')$$

where $x_I \equiv \left\{x \in X : x_i = \hat{x}_i \in H_i \quad \forall i \in I\right\}$. Here scenario s is – conditional on having fixed a certain $x_I \equiv (\hat{x}_i \in H_i)_{i \in I}$ – the exact equivalent of the scenario in Definition 1. When all dimensions $i \in I$ take exact values, expression (8') boils down into (8). More generally, criterion (8') relies on a two stage procedure. First, each hypothesis $h_r \cap d$ is decomposed into all its constituent "elementary hypotheses", defined as those that fix one exact value for each dimension in *I*.. For each elementary hypothesis (8'), the scenario maximizing the hypothesis' conditional probability is selected. Finally, (8') picks the elementary hypothesis that, with the respective maximizing scenario, has the highest conditional probability.[13] A solution to problem (8') always exists due to finiteness of the problem. This procedure generates a representation $s_r^1 \cap x_{I,r}^1 \cap d$ for hypothesis h_r which is the general counterpart of the outcome $s_r^1 \cap h_r \cap d$, obtained in the class of problems in (7). Accordingly, (8') yields a ranking $s_r^k \cap x_{I,r}^k$ among all possible representations of h_r that in turns ranks all elements in $h_r \cap d$ in terms of their order of recall. Formula (9) can be directly applied to calculate the local thinker's probabilistic assessment. Furthermore, in the case of exhaustive hypotheses in the general class (27) the local thinker's probabilistic assessment can be written as:

$$\Pr^L(h_t|d) = \frac{\left[\sum_{k=1}^{b} \Pr(s_t^k \cap x_{I,t}^k | h_t \cap d)\right] \Pr(h_t \cap d)}{\sum_{r=1}^{N} \left[\sum_{k=1}^{b} \Pr(s_r^k \cap x_{I,r}^k | h_r \cap d)\right] \Pr(h_r \cap d)} \quad (9'')$$

Expression (9'') is an immediate generalization of (9'). In particular, besides Proposition 1, which – as we shall see below is proved only for problems in (7) – all the results of the

[13] This assumption captures the idea that dimensions explicitly mentioned in the hypothesis are selected to maximize the probability of the latter. 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 representative, as expression (8) would always be equal to 1. Assumption (8') captures the intuitive idea that the agent also orders the representativeness of elements belonging to ranges explicitly mentioned in the hypothesis itself.

paper are generalized to hypotheses of kind (27), with the only caveat that in this case stereotype $s_r^k \cap h_r \cap d$ should be read as the intersection of the set of specific values chosen by the agent for representing h_r with the data and the chosen scenario, i.e. as $s_r^k \cap x_{I,r}^k \cap d$, where the latter term is the k'th ranked term according to objective (8').

Appendix 2: Proofs

Proof of Proposition 1. This proof restrict the analysis to the case where hypotheses h_1 and h_2 belong to class (7). Before proving the proposition, we identify one specific setting where hypotheses h_1, and $h_2 = \overline{h}_1$ have the same set of feasible scenarios. Let $X = \{0,1\}^K$ be the state space, generated by the product of *K*>2 binary dimensions (there is no loss of generality here as any finite state space can be represented this way). Focus on the class of problems where: i) the data *d* uniquely fix the value of *N*-1 dimensions, and ii) the agent assesses two hypotheses h_1, h_2 such that h_1 fixes the value of one dimension and h_2 fixes the other value of the same dimension, so that $h_2 = \overline{h}_1$. It is easy to check that in this case it is indeed true that $S_1 = S_2 = S$. Let us now prove Proposition 1 in this setting, starting with claim 1). Since h_2, h_1 each fix one value of the same binary dimension, Definition 1 applies and so does Assumption A2. Thus, since $h_2 = \overline{h}_1$, the representativeness ranking of $s \in S$ for h_1 follows $\Pr(h_1 | s \cap d) = \Pr(h_1 \cap d \cap s) / [\Pr(h_1 \cap d \cap s) + \Pr(h_2 \cap d \cap s)]$. The representativeness ranking of $s \in S$ for h_2 instead $\Pr(h_2 | s \cap d) = 1 - \Pr(h_1 \cap d | s)$. Evidently then, the representativeness levels of scenarios for the two hypotheses are perfectly inversely related, formally $s_1^k = s_2^{M-k+1}$ for $k = 1,\ldots,M$.

We now turn to claim 2.i). At any given $b < M$, h_1 is represented with scenarios $(s_1^k)_{k \le b}$, while h_2 is represented with $(s_1^{M+1-k})_{k \le b}$. 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 $b > b^*$. But then, this implies that (29) must hold for all $b > b^*$, including $b = M$, which is inconsistent with the fact that:

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

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

To see how in this case over-estimation of h_1 may be infinite, take by construction a $\pi(x)$ such that:

$$\Pr(s_1^k \cap h_1 \cap d) = \Pr(h_1 \cap d)\frac{1-\varepsilon^2}{1-\varepsilon^{2M}}\varepsilon^{2(k-1)},\ \Pr(s_1^k \cap h_2 \cap d) = \Pr(h_2 \cap d)\frac{1-\varepsilon}{1-\varepsilon^{M}}\varepsilon^{(k-1)}$$

, for all k≥1, where $0 < \varepsilon < 1$.

Then, for all $b \leq M$, we have that:

$$\sum_{k=1}^{b} \Pr(s_1^k \mid h_1 \cap d) = \frac{1-\varepsilon^{2b}}{1-\varepsilon^2}, \qquad \sum_{k=1}^{b} \Pr(s_1^{M+1-k} \mid h_2 \cap d) = \frac{\varepsilon^{M-b} - \varepsilon^{M}}{1-\varepsilon}$$

For $\varepsilon \to 0$ the extent of over-estimation is arbitrarily large for any $b < M$.

Finally, consider point 2.ii). If $\pi(x)$ is such that $\Pr(s_1^k \mid h_1 \cap d)$ decreases and $\Pr(s_1^k \mid h_2 \cap d)$ increase in k, the two hypotheses are represented with their most likely scenarios. Thus, the greatest over estimation of h_1 relative to h_2 is reached when $\Pr(s_1^1 \mid h_1 \cap d) = 1$ and $\Pr(s_1^M \mid h_2 \cap d) = 1/M$. That is, when h_1 is concentrated on its most likely scenario while the distribution of h_2 is fully dispersed among all scenarios. In this case, in which $\Pr(s_1^1 \mid h_1 \cap d) = 1/M$ and $\Pr(s_1^M \mid h_2 \cap d) = 1$, the agent under estimates the odds of h_1 by a factor of M/b.

Generalization of Proposition 2 to the Class of Problems in (27). This follows directly from the consideration that from (9) condition (18) directly translates into $\Pr(s_{1,2}^1 \cap x_{I_{1,2},1,2}^1) \geq \Pr(s_1^1 \cap x_{I_1,1}^1)$, which in turn implies that expression (19) becomes $\Pr(s_{1,2}^1 \cap x_{I_{1,2},1,2}^1 \mid h_1) \geq \Pr(s_1^1 \cap x_{I_1,1}^1 \mid h_1)$. This in turn implies that representation $s_1^1 \cap x_{I_1,1}^1$ must not be the most likely one for h_1 because it is also a representation of h_1.

Proof of Proposition 3. Before proceeding, it is important to bear in mind that in this proof we use the general representation rule (8') to encompass the case where the implicit disjunction h_1 specifies a range of values, as this more general case allows to explain the car mechanic experiment in a very simple setting. This implies that condition (24) now involves the full representation of a hypothesis, including the specification of the dimensions constrained by inequality in the hypothesis itself. Recall that in this case the expression $s_r^1 \cap h_r$ should be read as $s_r^1 \cap h_r(x_I^1)$ where $h_r(x_I^1)$ and s_r^1 maximize objective (8') To prove the proposition it is useful to first note that according to criterion

(8') representations follow a "revealed preference" logic: if the local thinker represents h_1 with $\left(x_I^1, s_1^1\right)$, then the agent will always use the same representation for a generic hypothesis $h_0 \subset h_1$ when $\left(x_I^1, s_1^1\right) \in h_0$ and, at the same time, s_1^1 is a feasible scenario for h_0, in the sense that h_0 and h_1 constrain the same set of dimensions *I*.

To see that in this case a representation for h_1 is also a representation for h_0, note that if the agent represents hypothesis h_1 with $\left(x_I^1, s_1^1\right)$ it must hold that:

$$\Pr\left[x_I^1 \middle| s_1^1 \cap d\right] > \Pr\left[x_I^1 \middle| s_1^{'} \cap d\right], \tag{31}$$

for every possible $x_I^{'}, s_1^{'}$. If however $\left(x_I^1, s_1^1\right)$ is *not* the representation of h_0, then (31) cannot hold. Suppose in fact that the representation of h_0 is equal to the element (x_I^*, s_0^*). Then, it must be true that:

$$\Pr\left[x_I^* \middle| s_0^* \cap d\right] > \Pr\left[x_I^1 \middle| s_1^1 \cap d\right]. \tag{32}$$

But condition (32) clearly violates (31) because representation (x_I^*, s_0^*) is also feasible for h_1 and would yield a higher value of criterion (8') than the original representation (x_I^1, s_1^1). Hence, if $(x_I^1, s_1^1) \in h_0$ and s_1^1 is a scenario for h_0, then h_0 has the same representation as h_1. But then, since by assumption s_1^1 is a scenario for either $h_{1,1}$ or $h_{1,2}$, it must be that (x_I^1, s_1^1) belongs to either of them. As a result, condition (24) holds and the disjunction fallacy obtains.

References

Cascells, Ward, Arno Schoenberger, and Thomas Graboys. 1978. "Interpretations of Physicians of Clinical Laboratory Results." *The New England Journal of Medicine*, 999-1001.

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.

Jehiel, Philippe. 2005. "Analogy-based Expectation Equilibrium," *Journal of Economic Theory* 123, 81-104.

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

Kahneman, Daniel, and Shane Frederick. 2005. "A Model of Heuristic Judgment," chapter 12 in K. Holyoake and R. Morrison, eds. *The Cambridge Handbook of Thinking and Reasoning*, Cambridge, UK: Cambridge University Press.

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.

_____. 1982. "Evidential Impact of Base-Rates," chapter 10 in D. Kahneman, P. Slovic, and A. Tversky, eds., *Judgement Under Uncertainty: Heuristics and Biases, Cambridge*, UKI Cambridge University Press.

Maya Bar-Hillel. 1982. "Studies of Representativeness," in D. Kahneman, P. Slovic, and A. Tversky, eds., *Judgment 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.

Osborne, Martin J. and Ariel Rubinstein. 1998. "Games with Procedurally Rational Players," *American Economic Review*, 88, 834-847.

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

Stewart, Neil, Nick Chater, and Gordon Brown, 2006. "Decision by sampling", *Cognitive Psychology*, 53, 1-26.

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.