Web Appendix 2: Proofs

Proof of Proposition 1. Proposition 1 is restricted to the case where hypotheses h_1 and h_2 belong to class (7). Note first that any finite state space can be represented as $X = \{0,1\}^K$ generated by the product of *K* binary dimensions. We assume that *K*>2 to allow for hypotheses, data and scenarios. If h_1 and h_2 have the same set of feasible scenarios ($S_1 = S_2$) then they necessarily fix the same set of dimensions ($I_1 = I_2$). Since dimensions are binary, it follows that $h_2 = \overline{h}_1$. For simplicity, focus on the class of problems where: i) the hypotheses h_1, h_2 fix the value of only one dimension and ii) the data *d* fix the value of *N*-1 other dimensions, *N*<*K*. The condition $S_1 = S_2 = S$ still holds.

To prove claim 1), apply Definition 1 and Assumption A2 to find that the representativeness of $s \in S$ for h_1 is equal to $\Pr(h_1 | s \cap d) = \Pr(h_1 \cap d \cap s)/[\Pr(h_1 \cap d \cap s) + \Pr(h_2 \cap d \cap s)]$. The representativeness of $s \in S$ for h_2 is equal to $\Pr(h_2 | s \cap d) = 1 - \Pr(h_1 | s \cap d)$. The representativeness of scenarios for the two hypotheses is thus perfectly inversely related, formally $s_1^k = s_2^{M-k+1}$ for *k* = 1,…,*M*.

Consider now claim 2.i). For any *b* < *M*, h_1 is represented with scenarios $\{s_1^k\}_{k \le b}$, while h_2 is represented with $\{s_1^{M+1-k}\}_{k \le b}$. From (9), the odds of h_1 are (weakly) over-estimated if and only if:

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

Suppose that $\Pr(s_1^k | h_1 \cap d)$ and $\Pr(s_1^k | h_2 \cap d)$ strictly decrease in *k*. It then follows that the above condition is met for every *b* < *M*. To establish a contradiction, suppose that for a certain $b^* < M$ the above condition is not met, that is

$$\sum_{k=1}^{b^*} \Pr(s_1^k | h_1 \cap d) < \sum_{k=1}^{b^*} \Pr(s_1^{M+1-k} | h_2 \cap d) \tag{26}$$

Then, for 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 since $\Pr(s_1^k | h_1 \cap d)$ and $\Pr(s_1^k | h_2 \cap d)$ are strictly decreasing 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 $b > b^*$. This implies that (26) holds for all $b > b^*$, including *b* = *M*, but this is inconsistent with the fact that

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

The same logic allows us to show that if $\Pr(s_1^k | h_1 \cap d)$ and $\Pr(s_1^k | h_2 \cap d)$ are strictly increasing in *k*, the odds of h_1 are (weakly) underestimated for any *b*<*M*.

To see how in the first case the overestimation of h_1 may be infinite, consider a probability distribution $\pi(x)$ 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 < ε < 1. Then, for all $b \leq M$, we have that:

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

Inserting these expressions into (11), we see that as ε → 0 the extent of overestimation becomes arbitrarily large for any $b < M$.

Finally, to prove claim 2.ii), recall that h_1 and h_2 are represented with scenarios s_1^1 and s_1^M respectively. 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 two hypotheses are represented with their most likely scenarios. Thus, the greatest overestimation of h_1 relative to h_2 is reached when h_1 is concentrated on its most likely scenario while the distribution of h_2 is fully dispersed among all scenarios, that is $\Pr(s_1^1 | h_1 \cap d) = 1$ and $\Pr(s_1^M | h_2 \cap d) = 1/M$. In this case, the agent overestimates the odds of h_1 by a factor of $\left(\sum_{k=1}^{b} 1/M\right)^{-1} = M/b$.

Proof of Proposition 2. To prove the proposition, we explicitly focus on hypotheses of the form in (7), but all of the results are easily extended to the case where hypotheses take the general form (7') by simply substituting h_i with $x_{I,i}^k$ when scenario s_i^k is used. The central part of the argument amounts to proving that if $s_i^1 \cap d \neq \phi$ and $s_i^1 \cap \overline{d} = \phi$ for all *i*, then stereotypes do not change. Formally, $s_i^1 \cap h_i = s_{i,d}^1 \cap h_i \cap d$ for all *i*, where $s_{i,d}^1$ is the most representative scenario after data *d* is provided. We prove this property by contradiction. If $s_i^1 \cap h_i \neq s_{i,d}^1 \cap h_i \cap d$ for some *i*, then it must also be the case that $s_i^1 \neq s_{i,d}^1 \cap d$ and therefore

$$\frac{\Pr(h_i \cap s_i^1)}{\Pr(\overline{h}_i \cap s_i^1)} = \frac{\Pr(h_i \cap s_i^1 \cap d)}{\Pr(\overline{h}_i \cap s_i^1 \cap d)} < \frac{\Pr(h_i \cap s_{i,d}^1 \cap d)}{\Pr(\overline{h}_i \cap s_{i,d}^1 \cap d)}. \tag{27}$$

Condition (27) follows from three considerations. First, since $s_i^1 \cap d \neq \phi$ and $s_i^1 \cap \overline{d} = \phi$ for all *i*, we have that $\Pr(h_i \cap s_i^1) = \Pr(h_i \cap s_i^1 \cap d)$, which implies the equality on the left hand side in (27). Second, since $s_i^1 \cap d \neq \phi$, then $s_i^1 \cap d$ contains a scenario for $h_i \cap d$ [this scenario is identified by the sub-vector *s* of elements in s_i^1 not fully pinned down by *d*]. This is because $s_i^1 \cap h_i \cap d$ identifies an element in X. Third, the scenario *s* identified in $s_i^1 \cap d$ must be less representative than $s_{i,d}^1$ because the latter is defined as the most representative scenario for

$h_i \cap d$. But then, since $s^1_{i,d} \cap d$ is also a scenario for h_i, the relationship between the first and third terms in condition (27) contradicts the fact that s^1_i is the most representative scenario for h_i. This proves that $s^1_i \cap h_i = s^1_{i,d} \cap h_i \cap d$, which directly implies that assessments do not change, upon provision of *d*, even if *d* is informative. If, in contrast, $s^1_i \cap d = \phi$ for some *i*, then the stereotype for the corresponding hypothesis must change. Then assessments can change even if the data is barely informative, as Section 5.3 and Appendix 3.A show. Here we show that the local thinker may even react to completely uninformative data. Consider the example below:

Data = d_1	s_1	s_2
h_1	ε_1	$\pi_1 - \varepsilon_1$
h_2	0	π_2

Data = d_2	s_1	s_2
h_1	0	π_1
h_2	ε_2	$\pi_2 - \varepsilon_2$

Table A2.1

The tables represent the distribution $\pi(x)$ on hypotheses h_1 and h_2 such that the data d_1, d_2 are completely uninformative (and ε_1, ε_2 are small positive numbers). When no data is provided, the local thinker represents h_1 with (s_1, d_1) and h_2 with (s_1, d_2), assessing $\Pr^L(h_1) = \varepsilon_1/(\varepsilon_1 + \varepsilon_2)$. After for instance d_1 is provided, the representation for h_1 does not change but the one for h_2 switches to (s_2, d_1). As a result, $\Pr^L(h_1|d_1) = \varepsilon_1/(\varepsilon_1 + \pi_2) << \Pr^L(h_1)$ even if the data is completely uninformative. This example is obviously extreme, but it gives an idea of the forces towards over-reaction in our model.

Generalization of Proposition 3 to the Class of Problems in (7'). Since b=1, each hypothesis h_i is represented by $x^1_{I_i,i} \cap s^1_i$, where $x^1_{I_i,i}, s^1_i$ satisfy (8'). Then condition (13) translates directly into $\Pr(s^1_{1,2} \cap x^1_{I_1,1} \cap x^1_{I_2,2}) \geq \Pr(s^1_1 \cap x^1_{I_1,1})$. Since both elements for which probabilities are computed in this condition are representations of h_1, we can rewrite this as $\Pr(s^1_{1,2} \cap x^1_{I_2,2} \cap x^1_{I_1,1}|h_1) \geq \Pr(s^1_1 \cap x^1_{I_1,1}|h_1)$. This in turn implies that representation $s^1_{1,2} \cap x^1_{I_1,1}$ must not be the most likely one for h_1, since $s^1_{1,2} \cap x^1_{I_1,1} \cap x^1_{I_2,2}$ is itself a more likely representation for h_1.

Proof of Proposition 4. We assume the implicit disjunction hypothesis $h_1 = h_{1,1} \cup h_{1,2}$ specifies a range of values, as this more general setting simplifies the analysis of the car mechanic experiment. In condition (15), the expression $s^1_r \cap h_r$ should be read as $s^1_r \cap h_r(x^1_I)$ where $h_r(x^1_I)$ and s^1_r satisfy (8'). Note that representations follow a "revealed preference" logic: if the local thinker represents h_1 with $\{x_I^1, s_1^1\}$, then he will always use the same representation for any hypothesis $h_0 \subset h_1$ as long as $x_I^1 \in h_0$ and 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 this, suppose that the representation of h_0 is equal to some other element $\{x_I^*, s_0^*\}$, so that:

$$\Pr\left(x_I^* \middle| s_0^* \cap d\right) > \Pr\left(x_I^1 \middle| s_1^1 \cap d\right).$$

But this leads to a contradiction, since $\{x_I^*, s_0^*\}$ would then be a representation of h_1 with higher conditional probability (8') than $\{x_I^1, s_1^1\}$. Continuing the proof, recall that by assumption s_1^1 is a scenario for either $h_{1,1}$ or $h_{1,2}$, or both. Therefore, $\{x_I^1, s_1^1\}$ is the representation of the hypotheses for which s_1^1 is a scenario. As a result, condition (15) holds and the disjunction fallacy follows.

Web Appendix 3. Additional Experiments

A. Insensitivity to Predictability

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 the 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.

To explore the consequences of local thinking on insensitivity to predictability, consider a local thinker who assesses the quality of a candidate based on the latter's job talk at a university department. The state space has 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 A3.1

Bad Talk (BT)	Inarticulate (I)	Articulate (A)
High Quality (H)	0.24	0.005
Low Quality (L)	0.25	0.005

Table A3.2

In tables A.1 and A.2, the quality of the talk is highly correlated with expressive ability, but the latter dimension is only barely informative of the candidate's quality. Still, the candidate's expressive ability is always representative of his quality, i.e. after listening to the talk the local thinker represents low quality candidates as inarticulate, and high quality ones as articulate. The tables are admittedly extreme, but they illustrate the point in the starkest manner. 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 this example a Bayesian would estimate Pr(H|GT)/Pr(L|GT) = 1.04 and Pr(H|BT)/Pr(L|BT) = 0.96 !!

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 A.1 and A.2 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.

B. Conjunction Fallacy in the Bjorn Borg Experiment

Suppose that a local thinker is given *d* = "Bjorn Borg is in the Wimbledon Final" and asked to assess Pr(Borg wins 1st set), Pr(Borg loses 1st set), Pr(Borg loses 1st and wins the match). The first hypothesis ensures exhaustivity, but it is not necessary to obtain the result. When prompted to assess these hypotheses, the agent fits an overall evaluation of Borg's game which can take two values: Borg loses the match (LM), Borg wins the match (WM). Suppose that the distribution of these characteristics is as follows:

Borg is in Wimbledon Final	Loses the Match (LM)	Wins the Match (WM)
Loses First Set (LS)	3/16	4/16
Wins First Set (WS)	2/16	7/16

Table A3.3

The Table above reports the actual fraction of each possible outcome observed in the 16 Grand Slam finals that Borg played between 1974 and 1981. The table reveals that the probability that Borg wins the final is large (equal to 11/16) irrespective of what happens in the first set, but losing the first set is relatively more likely if Borg loses the match (3 out of 5 rather than 4 out of 11). Crucially, the latter property implies that the agent represents the event WS with scenario WM and the event LS with scenario LM. By contrast, the hypothesis "Borg loses 1st set and wins the match" leaves no gap and is perfectly represented by (LS, WM). In this sate space it is easy to calculate that:

$$\frac{\Pr^L(LS,WM)}{\Pr^L(LS)} = \frac{\Pr(LS,WM)}{\Pr(LS,LM)} = \frac{4/16}{3/16} = \frac{4}{3} > 1 .$$

Thus, the conjunction rule is violated. Intuitively, the stereotypical condition in which the first set is lost is when the match is also lost. In computing Pr(LS) the local thinker overlooks the fact that Borg could lose the first set but actually win the match. The source of the conjunction fallacy here is that it is very unlikely for Borg to lose a Grand Slam (and thus Wimbledon final), even if he loses the first match.

C. Conjunction Fallacy Without Data Provision: Floods in California

Let the state space have the following three dimensions: the type of flood, which can either be severe (S) or disastrous (D), the cause of flood, which can either be an earthquake (E) or a rainstorm (R), and the location of the flood, which can either be California (C) or the rest of North America (NC). The distribution of outcomes is as follows:

S D	E	R
C	$(1-x)e_C$ xe_C	$r_C/2$ $r_C/2$
NC	$e_{NC}/2$ $e_{NC}/2$	$(1-z)r_{NC}$ zr_{NC}

Table A3.4

e_L and r_L capture the probabilities of an earthquake and a rainstorm in location $L = C, NC$, while $x > 1/2$ and $z > 1/2$ are respectively the share of earthquakes causing disastrous floods in California and of rainstorms causing disastrous floods in the rest of North America. Probabilities must add up to 1. Table B captures two features of a subject's beliefs: i) earthquakes are milder in the rest of North America than in California so that they cause fewer disastrous floods (only 1/2 of earthquakes cause disastrous floods in North America, x >1/2 earthquakes cause disastrous floods in California), and ii) rainstorms are milder in California than in the rest of North America so that they cause fewer disastrous floods (only 1/2 of rainstorms cause disastrous floods in California, z > 1/2 rainstorms cause disastrous floods in the rest of North America). We make the natural assumption that $z > x$, so that rainstorms are more likely to cause disastrous floods than earthquakes.

Table A.3 and equation (8) imply that a disastrous flood (D) is represented with scenario (R,NC), namely as a disastrous flood caused by a rainstorm in the rest of North America $\Pr(D|R,NC) = z \;>\; \Pr(D|E,C) = x \;>\; \Pr(D|R,C) = \Pr(D|E,NC) = 1/2$. The event "Disastrous flood caused by an earthquake in California" instead uniquely identifies the scenario (D, C, E). Given these representations, the assessed odds of (D,C,E) relative to (D) are:

$$\frac{\Pr^L(D)}{\Pr^L(D,C,E)} = \frac{\Pr(D,R,NC)}{\Pr(D,C,E)} = \frac{zr_{NC}}{xe_C}.$$

If the probability of disastrous earthquakes in California is sufficiently high relative to that of disastrous rainstorm in North America, (i.e., $xe_C > zr_{NC}$), the conjunction fallacy arises without data. Intuitively, although rainstorms mainly cause mild floods, they are a stereotypical cause of floods. Hence, disastrous floods are represented as being caused by rainstorms, even though agents hold the belief that earthquakes in California can be so severe as to cause more disastrous floods. The problem, though, is that agents retrieve this belief only if earthquakes and California are explicitly mentioned.