Online Supplemental Material to

Toward a Synthesis of Cognitive Biases:

how noisy information processing can bias human decision-making

Martin Hilbert

Annenberg School of Communication
University of Southern California

Appendix A: introductory analogy to memory-channel based decision-making schematizations

Appendix A gives an introduction to the logic of noise influenced memory-based decision making models and to the kinds of schematizations used to present them. It will help us to set the stage for what is to follow in the article.

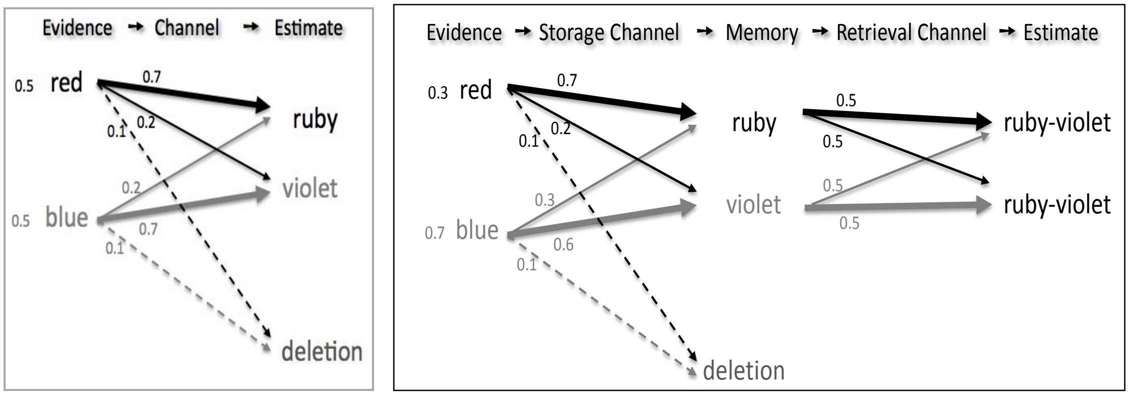
Let us start with a memory-based task. Suppose we would like to make a decision about the "redness" of a red object that we are given. Our strategy would consist in collecting and storing color traces which are purely red and once we are asked to judge redness, we would go to our storage room and pick up some of our prototypical red traces and compare their color with the object to be judged. In case our storage and retrieval processes would be perfect, we would be able to make a perfect judgment about the redness of the object. If this process is not perfect, we might erroneously end up with a sample that is not completely red and our judgment will be biased. The analysis of our storage and retrieval habits will show us the nature of this bias and be a first step in looking for strategies to minimize it.

In other words, when confronted with a memory-based decision-making task, the judge sends a cognitive probe to memory and compares it to existing memory traces. The content of what is found in memory will provide the judge with the answer to the decision problem. The process is not perfect. We will refer to the "confusion" and "mistakes" in this process as "noise". The bias of the judgment can be traced back to two possible sources: one is a biased sample in memory (which results from the noise in the storage channel); and the other one is biased sampling from memory (which results from the noise in the retrieval channel). The combination of the storage and retrieval channels constitutes the overall memory channel. We assume that the channel has certain properties that we would like to define.

In information theory it is customary to present these kinds of channels in a diagram similar to the ones shown in Figure A.1 (see Massey, 1998, Ch.4; Cover and Thomas, 2006, Ch.7). Figure A.1a essentially tells us that the noisy channel mixes blue and red input evidences. As long as the original still prevails, this will turn the red into a ruby and blue into violet. We depicted the noise with crossover arrows. The little numbers next to the arrows tell us about the respective transition probabilities involved in the process. Besides mixing evidence, we have to consider that not all input might make it. The effect is equal to deleting parts of our sample. We start out with equal amounts of red and blue [50% each]. 70% of each goes straight through the channel (we call this the identify transformation), 20% of each color is mixed with the other color (noise) and 10% of each is deleted.

Figure A.1b opens up the overall memory channel and shows that the overall memory channel actually consists of two different sub-channels. The noisy storage channel is followed by the noisy retrieval channel. In Figure A.1b, we start off with less red than blue [0.3, 0.7], and it is more likely to confuse blue with red [0.3], than red with blue [0.2]. The retrieval channel is in a special state of "highest uncertainty/entropy" (the uniform distribution), which leads to a homogeneous output estimate of ruby-violet, independent of the input evidence.

Figure A.1: Two first examples of the memory channel: (a) overall memory channel; (b) opened up into storage and retrieval subchannels



Source: author.

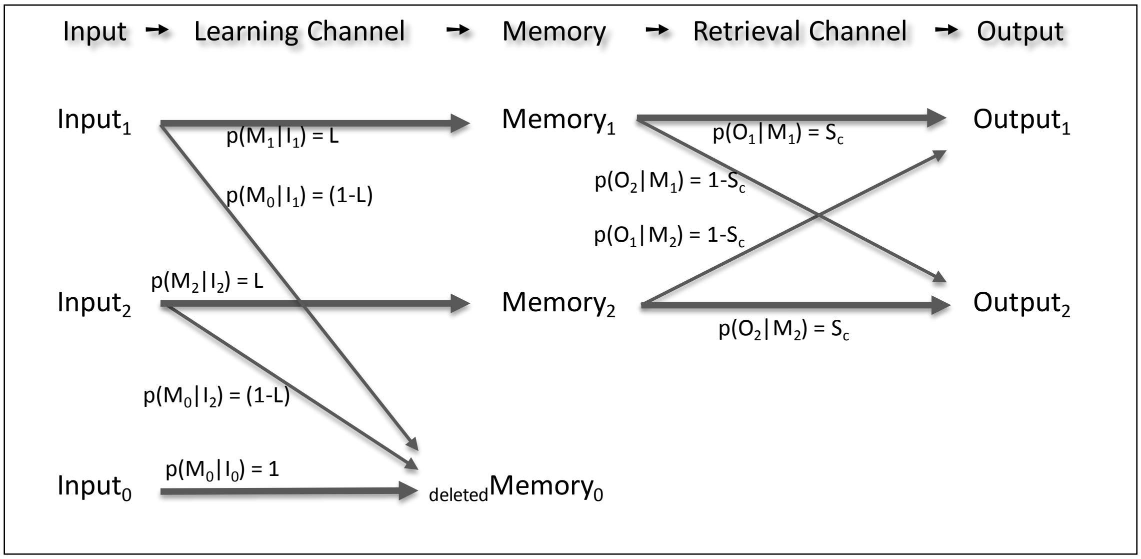
It is important to note that the intermediate memory step might be very short. Several perceptual tasks rely on sensory memory that corresponds approximately to the initial 200–500 milliseconds after an item is perceived. Therefore, the process might not appear as schematic as presented here (storing in memory, then sending a probe to memory, etc.), but rather as one process. Notwithstanding, without having anything impregnated in any kind of (whatsoever short and instable) memory, no perception could occur. Therefore, memory (of some kind) always makes part of any kind of judgment and decision-making.

Throughout the article, we will identify which kind of noise is requires in the overall channel (Figure A.1a) to replicate six cognitive biases, and which kind of noise is necessary in the retrieval channel (Figure A.1b) to replicate two additional biases.

Appendix B: the MINERVA-DM channel

Appendix B uses information-theoretic channel logic (see Appendix A) to discuss the essential properties of the MINERVA-DM channel (see Hintzman, 1988; Dougherty, et.al., 1999) (see Figure B.1). The goal is not to replicate the exact nature of the MINERVA-decision-making model, but to model its essential logic with the help of an information-theoretic channel presentation (see Appendix A; more formal in Massey, 1998, Ch.4; Cover and Thomas, 2006, Ch.7). Hintzman (1988) chooses a ternary input variable of -1, 0, +1, which is basically a binary alphabet, plus the possibility of deletion. The storage channel (in MINERVA called "learning channel") is implemented with what is known as a "Binary Erasure Channel" (BEC) in information theory. The retrieval channel is implemented with what is known as the "Binary Symmetric Channel" (BSC). The channel is symmetric because both identity transitions, and both noise transitions are equal, $p(O_1|M_1) = p(O_2|M_2) = S_c$, and $p(O_2|M_1) = p(O_1|M_2) = 1 - S_c$. Both are very special and important channels in information theory (Massey, 1998, Ch.4; Cover and Thomas, 2006, Ch.7). They are the simplest existing channels and their neat properties enable a straightforward analysis with nice results.

Figure B.1: Rough schematization of the MINERVA-DM model as a memory channel


Input → Learning Channel → Memory → Retrieval Channel → Output
Input_1
$p(M_1|I_1) = L$
Memory_1
$p(O_1|M_1) = S_c$
Output_1
$p(M_0|I_1) = (1\text{-}L)$
$p(O_2|M_1) = 1\text{-}S_c$
$p(O_1|M_2) = 1\text{-}S_c$
$p(M_2|I_2) = L$
Input_2
Memory_2
Output_2
$p(O_2|M_2) = S_c$
$p(M_0|I_2) = (1\text{-}L)$
Input_0
$p(M_0|I_0) = 1$
$_{\text{deleted}}\text{Memory}_0$

Source: author, based on the logic presented in Hintzman, 1988 and Dougherty, et.al, 1999.

The technical details of the specific implementation of MINERVA-DM are more involved than this simplified schematization. One aspect is that Hintzman (1988) chose a multi-trace memory model to implement MINERVA. He also chose not to apply the noise to an entire memory trace, but to its constituents, which he calls features. He uses a ternary code (-1, 0, +1) to represent the value of each feature, which make up the content of specific memories. Each of these features is passed through the channel, which has different probabilities of converting a -1 into a +1, and vice versa, or deleting it, which means converting it to 0. As the features change, the content of the memory trace change and it can even lead to the fact that the memory does not represent anymore what it originally meant to represent. The rate with which the content of the memory traces change, depends on the transition probabilities that convert the values of the features and on the criterion that defines when a code in a memory matches or not.

MINERVA also includes a particular matching process (which is not further justified by the authors). It basically replicates what is known as the Hamming distance between codewords in information theory. These particular specifications do not change the basic properties of the MINERVA-DM channel, which follows the logic of a BEC followed by a BSC: the storage/learning channel can delete input, and the retrieval channel has the possibility that "false friends" sneak into the final judgment.

Studying the logic of the MINERVA-DM channel, it becomes clear that the only way that the variation of L impacts the output is through a reduction of the sample size of traces in memory, by exactly [1-L], which leads to the well-known channel capacity of the BEC: C_{BEC}= L (see Massey, 1998, Ch.4; Cover and Thomas, 2006, Ch.7). The reason is that L is applied symmetrically to all inputs and that there is no crossover possibility in the storage/learning channel. This prediction was reconfirmed by the MINERVA-DM simulation results of Dougherty, et.al., 1999, shown in their Appendix C. As pointed out by them, the effect of smaller sample sizes is increased variability (the inverse of the law of the large numbers). On contrary, the retrieval channel, which is BSC, is sensible to variations in S_C (this is also in agreement with the simulation of Appendix C, Dougherty, et.al., 1999). The smaller S_C, the larger the crossover probability. The result is that both outputs are "more similar", i.e. they are closer to their "average", which is the uniform distribution (in this binary case: 0.5-0.5). Since the retrieval channel is BSC, its channel capacity is: C_{BSC}= 1-H(S_c) (see Massey, 1998, Ch.4; Cover and Thomas, 2006, Ch.7). MINERVA-DM applies noise of the same distribution to all input evidence. As we will show in Appendixes B and D, this requirement is not necessary to assure conservatism.

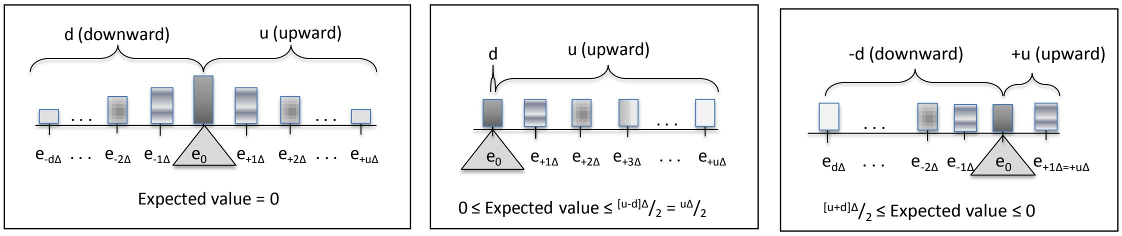
Appendix C: Effects of Properties N and S on a bounded noise distribution

This Appendix shows how Properties N and S lead to the fact that all mean estimates Ê, based on some concrete input evidence $\underline{e_i}$, (Expct.val.[Ê|$\underline{e_i}$]), must lie somewhere in the grey areas of Figure 3b. For likelihood/probability/frequency estimates, the interval variables e_i are replaced with probabilities. In this case, the exercise focuses directly at estimating the value Expct.val.[p(p(Ê)|p($\underline{e_i}$))]. For reasons of clarity of presentation we will treat both cases identically and refer to E instead of P(E).

We use a little trick and scale the identity transition e_i to 0: $\underline{e_0}$ = $\hat{e}_0$ = 0. We denote all estimates to the "positive" side with $\hat{e}_u$, u={0,1,2…u}, and estimates to the "negative" side of the identity transition with $\hat{e}_d$, d={0,1,2…d}. We stick to the assumption of a one-dimensional equidistant interval scale, which results in: $\hat{e}_u = \hat{e}_0 + u\Delta$; $\hat{e}_d = \hat{e}_0 - d\Delta$. The result looks like Figure Ca. Property N assures that none of the "weights" can be larger than the value assigned to e_0 (identity transition). Property N assures that the weights get smaller the further they are away from the identity value, and Property S demands that the weights are symmetrical around the identity value. The total number of possible values is: n = u+d+1, whereas +1 counts for the identity value at 0 (note that we do not consider the forgetting / inaccessibility option here. If we would, the number of possible values would be n+1).

Visually the logic of the proof can be seen when playing around with Figures C. When moving the balance triangle all the way to the negative extreme (d ≤ u), the minimum value is 0 and the maximum positive value can be achieved by placing the highest possible weight on the largest possible numbers (Figure Cb). Considering the restrictions of Property N, the uniform distribution achieves this maximum value (in this case: $^{u\Delta}/_{2}$). In Figure Cc, d ≥ u, and since d is preceded by a minus sign, the expected value can only be negative, with: $^{[u-d]\Delta}/_{2}$ ≤ Expected value ≤ 0.

Figure C: (a) representation of an evidence at the middle of the possible scale in the task; (b) representation of evidence at the lowest possible value of the scale; (c) representation of evidence one step from the highest possible value on the scale.


d (downward)
u (upward)
$e_{-d\Delta}$ . . . $e_{-2\Delta}$ $e_{-1\Delta}$ e_0 $e_{+1\Delta}$ $e_{+2\Delta}$ . . . $e_{+u\Delta}$
Expected value = 0
d
u (upward)
e_0 $e_{+1\Delta}$ $e_{+2\Delta}$ $e_{+3\Delta}$ . . . $e_{+u\Delta}$
0 ≤ Expected value ≤ $^{[u-d]\Delta}/_{2} = {}^{u\Delta}/_{2}$
-d (downward)
+u (upward)
$e_{d\Delta}$ . . . $e_{-2\Delta}$ $e_{-1\Delta}$ e_0 $e_{+1\Delta=+u\Delta}$
$^{[u+d]\Delta}/_{2}$ ≤ Expected value ≤ 0

Source: author.

In a more formal proof we first define the limits of the possible expected values (note that EV without hat and no underline, refers to "Expected Value", not to be confused with Ê: estimation; and E: evidence):

If d ≤ u, then:

$$0 \le EV\left[\hat{E}\middle|\underline{e_0}\right] \le \sum_{i=0}^{n}\hat{e}_i/n = \left[\sum_{d=0}^{d}\hat{e}_d + \sum_{u=0}^{u}\hat{e}_u\right]/n =$$

$$\left[\sum_{d=0}^{d}\hat{e}_0 - d\Delta + \sum_{u=0}^{u}\hat{e}_0 + u\Delta\right]/n =$$

$$= \Delta\left[-\sum_{d=0}^{d} d + \sum_{u=0}^{u} u\right]/n = \Delta\left[-\frac{d(d+1)}{2} + \frac{u(u+1)}{2}\right]/[u+d+1] =$$

$$= \Delta\left[\frac{-d^2-d+u^2+u}{2}\right]/[u+d+1] = \frac{[u-d]\Delta}{2} = \frac{\hat{e}_u+\hat{e}_d}{2} = its\ midrange\ point\ m$$

If d ≥ u, then:

$$0 \ge EV\left[\hat{E}\middle|\underline{e_0}\right] \ge \sum_{i=1}^{n}\hat{e}_i/n = \cdots = \frac{[u-d]\Delta}{2}$$

We now reformulate the EV[Ê|$\underline{e_0}$], following Properties S and N:

$$EV\left[\hat{E}\middle|\underline{e_0}\right] = \sum_{j=-d}^{u} p(\hat{e}_j \,|\, \underline{e_0}) \times \hat{e}_j = \sum_{d=0}^{d} p(\hat{e}_d \,|\, \underline{e_0}) \times \hat{e}_d + \sum_{u=0}^{u} p(\hat{e}_u \,|\, \underline{e_0}) \times \hat{e}_u =$$

$$= \sum_{d=0}^{d} p(\hat{e}_d \,|\, \underline{e_0}) \times (\hat{e}_0 - d\Delta) + \sum_{u=0}^{u} p(\hat{e}_u \,|\, \underline{e_0}) \times (\hat{e}_0 + u\Delta) =$$

If d ≤ u (group all possible symmetric noises under the same sum and cancel them out):

$$EV\left[\hat{E}\middle|\underline{e_0}\right] = \sum_{k=0}^{d} p(\hat{e}_k \,|\, \underline{e_0}) \times (\hat{e}_0 - k\Delta + \hat{e}_0 + k\Delta) + \sum_{u=d+1}^{u} p(\hat{e}_u \,|\, \underline{e_0}) \times (\hat{e}_0 + \Delta u) = \sum_{u=d+1}^{u} p(\hat{e}_u \,|\, \underline{e_0}) \times \Delta u =$$

≥ 0 (its minimum), achieved if $p(\hat{e}_u|\underline{e_0})=0$, for all d<u;

≤ with its maximum if $P(\hat{E}|\underline{e_0}) = {}^{1}/_{n}$ (limited by Property N to the uniform distribution, which puts most weight on the positive extremes), at:

$$\leq \frac{1}{n}\sum_{u=d+1}^{u} \Delta u = \frac{1}{u+d+1}\Delta \sum_{u=d+1}^{d+(u-d)} u = \frac{\Delta}{u+d+1}\,[(d+1)+(d+2)+\cdots+(d+(u-d)] =$$

$$= \frac{\Delta}{u+d+1}\left[d(u-d) + \sum_{t=1}^{u-d} t\right] = \frac{\Delta}{u+d+1}\left[\frac{2d(u-d)}{2} + \frac{(u-d)[(u-d)+1]}{2}\right] =$$

$$= \frac{\Delta}{u+d+1} \times \frac{-d^2+u^2+u-d}{2} = \frac{(u-d)\Delta}{2}$$

=> if d ≤u, then 0 ≤ $EV[\hat{E}|\underline{e_0}] \leq {}^{[(u-d)\Delta]}/_{2} = [\hat{e}_u+\hat{e}_d\,]/_{2}$= its midrange point m, see Figure 3b.

If d ≥ u (group all possible symmetric noises under the same sum and cancel them out):

$$EV\left[\hat{E}\middle|\underline{e_0}\right] = \sum_{k=0}^{u} p(\hat{e}_k \,|\, \underline{e_0}) \times (\hat{e}_0 - k\Delta + \hat{e}_0 + k\Delta) + \sum_{d=u+1}^{d} p(\hat{e}_d \,|\, \underline{e_0}) \times (\hat{e}_0 - \Delta d) = \sum_{d=u+1}^{d} p(\hat{e}_d \,|\, \underline{e_0}) \times (-\Delta d) =$$

≤ 0 (its maximum), if $p(\hat{e}_d|\underline{e_0})=0$, for all d>u;

≥ with its minimum if $P(\hat{E}|\underline{e_0}) = {}^{1}/_{n}$ (uniform, limited by Property N), at:

$$\geq \frac{1}{n}\sum_{d=u+1}^{d}(-\Delta d) = \frac{-\Delta}{d+u+1}\sum_{d=u+1}^{u+(d-u)} d = \mathit{following\ the\ same\ steps\ as\ above} = \frac{(u-d)\Delta}{2}$$

=> if d ≥ u, then 0 ≥ $EV[\hat{E}|\underline{e_0}] \geq {}^{[(u-d)\Delta]}/_{2} = [\hat{e}_u+\hat{e}_d\,]/_{2}$= its midrange point mr, see Figure 3b.

Appendix D: Fitting the Gaussian channel

This Appendix shows how to convert a Gaussian channel into discrete transition probabilities and how to fit it to empirical finding. Usually, the most straightforward way of fitting normal noise to empirical findings is to set up a program (for example in MATLAB) that minimizes the distance between the empirical transition matrix and the modeled transition matrix, which is defined by the cutoff criteria between the variables, and the mean and variables of the normal distribution. Least squares can be used.

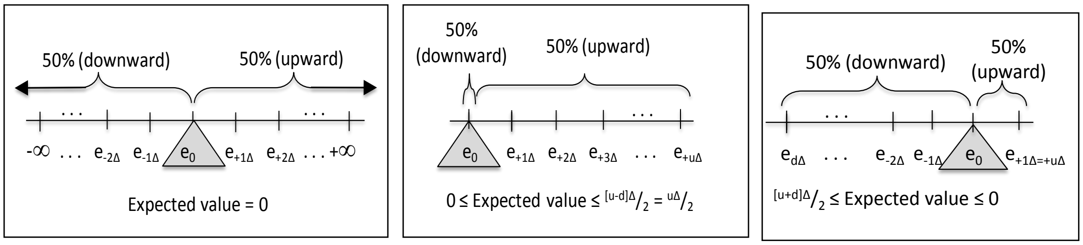
In the following I will go "manually" through the simple process of the ternary exercise of Figure 10. The problem that we face is that we have two degrees of freedom to work with when fitting the curves to the empirical data: the mean and variance of the normal distribution. However, since we suppose an equidistant interval scale (in this case $\Delta \approx 3.83\sigma$), the means of all three normal curves are defined once two means are chosen. The remaining degrees of freedom stem from the adjustable variances. This implies that it is not possible to perfectly fit the three curves to the six degrees of freedom of the empirical finding. We would not have had this problem in a binary decision-making task.

We start by arbitrarily defining that $\underline{e}_1$ fits the standard normal curve with $\mu=0$ and $\sigma=1$. We know that the identity transition of $\underline{e}_1$ is 0.971 and can therefore use the inverse of the cumulative normal distribution to identify the value x: $\Phi(1.896) \approx 0.971$, where $\Phi(x)$ is the cumulative standard normal function. We then have the freedom of choice for the mean and variance of $\underline{e}_2$ (defined by the variables of the normal probability density function: $f(x) = {}^{1}/_{\sqrt{(2\pi)}\sigma}\ e^{-(x-\mu)^2/2\sigma^2}$), as well as for the variance of $\underline{e}_3$ (once we define the mean of $\underline{e}_2$, the equidistance requirement determines the mean of $\underline{e}_3$). Instead of fitting all estimates as good as possible, I took the deliberate decision to "sacrifice" the fit of the noise-transitions $p(\hat{e}_2|\underline{e}_1)$ and $p(\hat{e}_3|\underline{e}_1)$, since they are very small. It turns out a mean of $\mu_2=3.83$ and variance of $\sigma_2=1.33$ for $\underline{e}_2$ and a variance of $\sigma_3=2.92$ for $\underline{e}_3$ fit to model the remaining transition probabilities of those curves. As expected, the "cost" paid by the mismatch of $\underline{e}_1$ is not very large.

Appendix E: Effects of Properties S and U on an unbounded noise distribution

This Appendix shows how Properties S and U lead to the fact that all average estimates Ê, based on some concrete input evidence $\underline{e}_i$, (Expct.val.[Ê|$\underline{e}_i$]), must lie somewhere in the grey areas of Figure 3b. As in Appendix C, we will refer to exercises that focus on estimating absolute numbers, E, but the same argument holds for estimations of discretized probabilities P(E). The basic logic of the effect of Property U can be seen when looking at what happens when we add the overshooting noise to the extremes of a symmetric distribution. A symmetric distribution around $\underline{e}_i$ has expected value = $\underline{e}_i$. We can normalize Expct.val.[$\underline{E}$] = 0 (see Figure Ea). When the valid scale is limited on the left side at the identity (Figure Eb), and the weight of the (formerly) negative values is added to the left-extreme value 0, the expected value ≤ ${}^{u\Delta}/_2$. When the scale is limited to the right (Figure Ec): Expected value ≥ ${}^{[u-d]\Delta}/_2$. Following this logic results in the fact that the subjective estimates Ê, must lie somewhere within the grey-shaded areas in Figure 3b.

Figure E: (a) normalized around 0; (b) left overshoot added to negative extreme; (c) right overshoot added to positive extreme.


50% (downward)
50% (upward)
-∞ . . . e-2Δ e-1Δ e0 e+1Δ e+2Δ . . . +∞
Expected value = 0
50% (downward)
50% (upward)
e0 e+1Δ e+2Δ e+3Δ . . . e+uΔ
0 ≤ Expected value ≤ [u-d]Δ/2 = uΔ/2
50% (downward)
50% (upward)
edΔ . . . e-2Δ e-1Δ e0 e+1Δ=+uΔ
[u+d]Δ/2 ≤ Expected value ≤ 0

Source: author.

The formal proof follows the same notations as in Appendix C, with the addition that the unbounded noise is defined by v = {-∞… -1, 0, +1, … +∞}. As in Appendix C have to proof that:

If d ≤ u then $0 \leq E[\hat{E}|\underline{e_0}] \leq {}^{[(u\text{-}d)\Delta]}/_2$.

If d ≥ u then $0 \geq E[\hat{E}|\underline{e_0}] \geq {}^{[(u\text{-}d)\Delta]}/_2$.

We now reformulate the EV[Ê|$\underline{e_0}$], following Properties S and U:

$$EV\left[\hat{E}\middle|\underline{e_0}\right] = \sum_{j=-d}^{u} p(\hat{e}_j\,|\,\underline{e_0}) \times \hat{e}_j + \sum_{v<-d}^{-\infty} p(\hat{e}_v\,|\,\underline{e_0}) \times \hat{e}_d + \sum_{v>u}^{+\infty} p(\hat{e}_v\,|\,\underline{e_0}) \times \hat{e}_u = \sum_{d=0}^{d} p(\hat{e}_d\,|\,\underline{e_0}) \times (\hat{e}_0 - d\Delta) + \sum_{u=0}^{u} p(\hat{e}_u\,|\,\underline{e_0}) \times (\hat{e}_0 + u\Delta) + \sum_{v<-d}^{-\infty} p(\hat{e}_v\,|\,\underline{e_0}) \times \hat{e}_d + \sum_{v>u}^{+\infty} p(\hat{e}_v\,|\,\underline{e_0}) \times \hat{e}_u =$$

If d ≤ u (group all possible symmetric noises under the same sum and cancel them out):

$$= \sum_{k=0}^{d} p(\hat{e}_k\,|\,\underline{e_0}) \times (\hat{e}_0 - k\Delta + \hat{e}_0 + k\Delta) + \sum_{u=d+1}^{u} p(\hat{e}_u\,|\,\underline{e_0}) \times (\hat{e}_0 - d\Delta + \hat{e}_0 + \Delta u) + \sum_{v>u}^{+\infty} p(\hat{e}_v\,|\,\underline{e_0}) \times (\hat{e}_0 - v + \hat{e}_0 + v) = \sum_{u=d+1}^{u} p(\hat{e}_u\,|\,\underline{e_0}) \times (u-d)\Delta$$

Whereas d is a constant ≤ u and u are positive integers.

≥ 0 (its minimum) if p($\hat{e}_u$|$\underline{e_0}$)=0, for all d≤u;

≤ its maximum at: 0.5 x (u-d)Δ = ${}^{(u\text{-}d)\Delta}/_2$ = [$\hat{e}_u$+$\hat{e}_d$]/2= its midrange point, see Figure 3b (given symmetry of unbounded distribution, the maximum possible weight on the positive u-side is 0.5).

If d ≥ u (group all possible symmetric noises under the same sum and cancel them out):

$$= \mathit{following\ t□e\ same\ steps\ as\ above} = \sum_{d=u+1}^{d} p(\hat{e}_d\,|\,\underline{e_0}) \times (u-d)\Delta$$

Whereas u is a constant ≤ d and d positive integers.

≤ 0 (its maximum) if p($\hat{e}_u$|$\underline{e_0}$)=0, for all d≥u;

≥ its minimum at: 0.5 x (u-d)Δ = ${}^{(u\text{-}d)\Delta}/_2$ = [$\hat{e}_u$+$\hat{e}_d$]/2= its midrange point, see Figure 7 (given symmetry of unbounded distribution, the maximum possible weight on the negative d-side is 0.5).

Appendix F: Effects of Properties D and N

We proof that a single-peak unimodal (Property N) noise distribution that has a doubly stochastic transition matrix (Property D) results in regressive behavior for any kind of input distribution (conservatism). We start with the reformulation of our conservatism requirement, equation (III):

$0 \leq \mathrm{cov}(\underline{\mathrm{E}},\hat{\mathrm{E}}) \leq \mathrm{Var}(\underline{\mathrm{E}})$

$0 \leq EV[\underline{\mathrm{E}}\hat{\mathrm{E}}] - EV[\underline{\mathrm{E}}] * EV[\hat{\mathrm{E}}] \leq EV[\underline{\mathrm{E}}^2] - (EV[\underline{\mathrm{E}}])^2$; (**Error! Bookmark not defined.**)

For n-ary decision-making tasks, scale to:

$EV[\underline{\mathrm{E}}] = \sum_{k=1}^{n} p(\underline{e}_k)\, \underline{e}_k \;= 0$; whereas $\underline{e}_k$ are positive numbers (in case of likelihood estimates $\sum_{k=1}^{n} \underline{e}_k \;= 1$ and $p(\underline{e}_k)$ is uniform with $p(\underline{e}_k) = \frac{1}{n}$ (like in Hockley's exercise, or, for likelihood estimates, one can imagine that likelihoods are represented by n memory traces, each representing the likelihood with its respective value $\underline{e}_k$).

=> $0 \leq EV[\underline{\mathrm{E}}\hat{\mathrm{E}}] \leq EV[\underline{\mathrm{E}}^2]$;

$0 \leq \sum_{k=1}^{n} \sum_{j=1}^{n} p(\underline{e}_k \hat{e}_j)\, \underline{e}_k \hat{e}_j \leq \sum_{k=1}^{n} p(\underline{e}_k)\, \underline{e}_k^2$;

$0 \leq \sum_{k=1}^{n} \sum_{j=1}^{n} p(\underline{e}_k)\, p(\hat{e}_j|\underline{e}_k)\, \underline{e}_k \hat{e}_j \leq \sum_{k=1}^{n} p(\underline{e}_k)\, \underline{e}_k^2$; multiplied with n;

$0 \leq \sum_{k=1}^{n} \sum_{j=1}^{n} p(\hat{e}_j|\underline{e}_k)\, \underline{e}_k \hat{e}_j \leq \sum_{k=1}^{n} \underline{e}_k^2$; whereas $p(\hat{e}_j|\underline{e}_k)$are the doubly stochastic weights of the transition matrix, with $\sum_{j=1}^{n} p(\hat{e}_j|\underline{e}_k) = 1$ and $\sum_{k=1}^{n} p(\hat{e}_j|\underline{e}_k) = 1$.

First, we focus on the left side of the inequality, for which we will show that Properties N and D assure that the resulting correlation cannot be negative:

$0 \leq \sum_{k=1}^{n} \sum_{j=1}^{n} p(\hat{e}_j|\underline{e}_k)\, \underline{e}_k \hat{e}_j = \sum_{k=1}^{n} \underline{e}_k \sum_{j=1}^{n} \hat{e}_j\, [p(\hat{e}_k|\underline{e}_k) - d_j]$; whereas d_j consists of positive numbers which represent how much smaller the noise is than the identity transition $p(\hat{e}_k|\underline{e}_k)$. Note that, according to Property N, d_j = 0 at $d_{j=k}$, and increases with j being more distant from k.

=> $\sum_{k=1}^{n} \underline{e}_k \left[\sum_{j=1}^{n} \hat{e}_j\, p(\hat{e}_k|\underline{e}_k) - \sum_{j=1}^{n} \hat{e}_j\, d_j\right] = \sum_{k=1}^{n} \underline{e}_k \sum_{j=1}^{n} \hat{e}_j\; (-d_j)$;

Note that there are positive and negative values of $\hat{e}_j$, since $EV[\underline{\mathrm{E}}] = EV[\hat{\mathrm{E}}] = 0$, therefore let $\sum_{j=1}^{n} \hat{e}_j = \left[\sum_{j-=-j}^{0} \hat{e}_{j-}\right] + \left[\sum_{j+=0}^{+j} \hat{e}_{j+}\right]$, with |-j|+|+j|=n, whereas $\hat{e}_{j-}$ denotes all negative values of $\hat{e}_j$, and $\hat{e}_{j+}$ stands for all positive values of $\hat{e}_j$. Since both parts have the equal weight, it is possible to rearrange both sums and organize them in according to equal distributions ("weigh them against each other", which we call w), with – $[\sum_{j-=-j}^{0} \hat{e}_{j-}] = [\sum_{j+=0}^{+j} \hat{e}_{j+}] \;=\; \sum_{m=1}^{m} w_m$, with w always being positive. Let d_{m-} correspond to d_j of $\hat{e}_{j-}$ and d_{m+} correspond to d_j of $\hat{e}_{j+}$:

$=> \sum_{k=1}^{n} \underline{e}_k \, [\sum_{j-=-j}^{0} \hat{e}_{j-}\left(-d_{j-}\right) + \sum_{j+=0}^{+j} \hat{e}_{j+}\left(-d_{j+}\right)] \; = \sum_{k=1}^{n} \underline{e}_k \, [\sum_{m=1}^{n} w_m(d_{m-}) +$
$\sum_{m=1}^{n} w_m(-d_{m+})] \; = \sum_{k=1}^{n} \underline{e}_k \, \sum_{m=1}^{n} w_m(d_{m-} - d_{m+}) = \sum_{k=1}^{n} \underline{e}_k \, \sum_{m=1}^{n} w_m\left(d_{+/-}\right).$

For negative $\underline{e}_k : d_{+/-}$ is also negative for all w_m with m≥k (given Property N), but not necessarily for w_m with m<k (since noise is not symmetric around identity m=k). Likewise, for positive $\underline{e}_k : d_{+/-}$ is also positive for all w_m with m≤k, but not necessarily for w_m with m>k. However, when rearranging to $\sum_{m=1}^{n} w_m \, \sum_{k=1}^{n} \underline{e}_k \left(d_{+/-}\right)$, we can see (since $EV\left[\underline{\mathrm{E}}\right] = \sum_{k=1}^{n} \underline{e}_k \; = 0$), that it is impossible that these eventualities drag the second sum into the negative

$=> \; \sum_{m=1}^{n} w_m \, \sum_{k=1}^{n} \underline{e}_k \, d_{+/-} \geq 0.$

This shows that the correlation cannot be negative ($0 \leq EV\left[\underline{\mathrm{E}}\hat{\mathrm{E}}\right]$). The right side of our initial inequality from equation (III) can be shown with similar reformulations, but actually, this proof and its insight are not new. It is very well known in information theory that a doubly stochastic transition matrix converts the channel input in a way that the output is overall closer to its mean ("stochastic mixing increases entropy") (see Cover and Thomas, 2006, Ch.4, p. 88, Exercise 4.1). We have defined conservatism as the output being "closer to the mean" than the input (see equation (II), formulated in variance). The new part is, that in our case, we claim that conservatism also implies a positive correlation between input and output (out estimates have "more to do with the evidence than they don't"). Property N assures this, as shown above.