

Only Fools Avoid Hindsight Bias: Appendix

Kevin Dorst
kmdorst@mit.edu
MIT
July 2024

Appendix 1: A model of hindsight bias

The point of this model is to show that there’s nothing fishy about your prior P being an object of uncertainty, and give a concrete model in which hindsight bias is rational.

We do this with the structure of **probability frames**—generalizations of Kripke frames from epistemic logic to probabilistic settings (Kripke 1963; Hintikka 1962; Williamson 2008; Dorst et al. 2021), and variants of Harsanyi type spaces (Harsanyi 1967).

A probability frame $(W, P, @)$ is a triple of a set of worlds (outcome space) W , a designated actual world $@$, and a function P from worlds w to probability distributions P_w defined over all subsets of W . P can be thought of as a variable for your prior probability distribution, or as a *description*, ‘Your prior distribution, whatever it is’. Since it can pick out different functions at different worlds, you can be uncertain what probability function you have. Meanwhile, P_w is a rigid designator for the particular probability distribution you have at world w . $P_@$ is your actual distribution—the one you have at the actual world.

We use P to define events about your probabilities: for any proposition $q \subseteq W$ and number $a \in \mathbb{R}$, $\langle P(q) = a \rangle := \{w \in W : P_w(q) = a\}$. This event will get assigned probabilities by each of the P_w , just like any other event.

Given this, let’s specify a frame in which $P_@$ commits hindsight bias. We can do this with a stochastic matrix: an $N \times N$ matrix in which row i column j indicates the probability that world i assigns to world j , $P_i(j)$. Thus row i is the probability vector P_i , in which the j th entry is the probability that P_i assigns to the j th world. (This is a familiar way to specify a Markov chain, which is formally identical to a probability frame but has a different interpretation.)

Let e be the proposition that *Biden is too old*. Suppose you start out uncertain of this proposition, but also uncertain about how uncertain you are. In particular,

suppose that *in fact* you have a middling 50%-credence in this, but you leave open that you might have a low 40% or high 60% instead. There are then 6 relevant possibilities: e_l where he's too old and you have a low (40%) credence in it, \bar{e}_l where he's not too old and you have a low credence that he is; e_m where he's too old and you have a middling (50%) credence he is, etc. There are some choice points here, but here's simple probability frame modeling this situation:

$$\left(\begin{array}{c|cccccc} & e_l & \bar{e}_l & e_m & \bar{e}_m & e_h & \bar{e}_h \\ \hline e_l & .05 & .45 & .125 & .125 & .225 & .025 \\ \bar{e}_l & .05 & .45 & .125 & .125 & .225 & .025 \\ \mathbf{e_m} & \mathbf{.025} & \mathbf{.225} & \mathbf{.25} & \mathbf{.25} & \mathbf{.225} & \mathbf{.025} \\ \bar{e}_m & .025 & .225 & .25 & .25 & .225 & .025 \\ e_h & .025 & .225 & .125 & .125 & .45 & .05 \\ \bar{e}_h & .025 & .225 & .125 & .125 & .45 & .05 \end{array} \right)$$

In this frame, the actual world is the bolded row e_m : so Biden is too old and your prior is 50% in this, since, summing across the e -worlds, $.025 + .25 + .225 = 0.5$. Thus $P_{\textcircled{a}}(e) = 0.5$. But you leave open that you're at an l -world (e_l or \bar{e}_l) or an h -world—in fact, you assign 25%-credence to each of those two events. And at those worlds, $P(e)$ is different. For $w \in \{e_l, \bar{e}_l\}$, $P_w(e) = .05 + .125 + .225 = 0.4$, while for $x \in \{e_h, \bar{e}_h\}$, $P_x(e) = .025 + .125 + .45 = 0.6$. Thus at the actual world you leave open that you're 40%- or 60%-confident of e : $P_{\textcircled{a}}(e) = 0.5$, but $P_{\textcircled{a}}(\langle P(e) = 0.4 \rangle) = P_{\textcircled{a}}(\{e_l, \bar{e}_l\}) = .025 + .225 = 0.25$, and likewise $P_{\textcircled{a}}(\langle P(e) = 0.6 \rangle) = P_{\textcircled{a}}(\{e_h, \bar{e}_h\}) = .225 + .025 = 0.25$. This is a mathematically coherent model of higher-order uncertainty.

Why does it exhibit hindsight bias? Notice that $P_{\textcircled{a}}$ starts out with the distribution $(0.25, 0.5, 0.25)$ over the three possible credences P might assign to e , i.e. 0.4, 0.5, and 0.6. That means its initial expectation for its prior credence is $\mathbb{E}_{P_{\textcircled{a}}}(P(e)) = 0.25(0.4) + 0.5(0.5) + 0.25(0.6) = 0.5$.

What happens when $P_{\textcircled{a}}$ learns e ? It updates to $P_{\textcircled{a}}(\cdot|e)$ by zeroing out \bar{e} -possibilities and re-normalizing. In particular, it shifts to the distribution $\left(\begin{array}{cccccc} e_l & \bar{e}_l & e_m & \bar{e}_m & e_h & \bar{e}_h \\ \hline .05 & 0 & .5 & 0 & .45 & 0 \end{array} \right)$. And notice that this distribution has shifted away from $\langle P(e) = 0.4 \rangle$ -possibilities and toward $\langle P(e) = 0.6 \rangle$ -possibilities: it now assigns 5%-credence to the former and 45%-credence to the latter (rather than the original 25%-25% each). As a result, $P_{\textcircled{a}}$'s posterior expectation for its prior credence in e is $\mathbb{E}_{P_{\textcircled{a}}}(P(e)|e) = .05(0.4) + .5(0.5) + .45(0.6) = 0.54$, which is greater than its prior estimate for its prior of $\mathbb{E}_{P_{\textcircled{a}}}(P(e)) = 0.5$.

If we want to update the entire frame on the true cell of the partition $\{e, \bar{e}\}$, we

get the following frame:

$$\left(\begin{array}{c|cccccc} & e_l & \bar{e}_l & e_m & \bar{e}_m & e_h & \bar{e}_h \\ \hline e_l & .125 & 0 & .3125 & 0 & .5625 & 0 \\ \bar{e}_l & 0 & .75 & 0 & .208333 & 0 & .0416667 \\ e_m & \mathbf{.05} & \mathbf{0} & \mathbf{.5} & \mathbf{0} & \mathbf{.45} & \mathbf{0} \\ \bar{e}_m & 0 & .45 & 0 & .5 & 0 & .05 \\ e_h & .0416667 & 0 & .208333 & 0 & .75 & 0 \\ \bar{e}_h & 0 & .5625 & 0 & .3125 & 0 & .125 \end{array} \right)$$

As can be checked, each world w , the distribution P_w exhibits hindsight bias: at worlds where it learns e , it increases its estimate for its prior in e ; at worlds where it learns \bar{e} , it decreases its estimate for its prior in e .

Appendix 2: Proof that $Cov[P(e), \mathbb{1}_e] > 0 \Leftrightarrow$ hindsight bias

Here's the fact to be proven:

Fact. If $Cov_{P_{\mathbb{Q}}}[P(e), \mathbb{1}_e] > 0$ if and only if $\mathbb{E}_{P_{\mathbb{Q}}}(P(e)|e) > \mathbb{E}_{P_{\mathbb{Q}}}(P(e))$.

Recall the definitions:

- $Cov_{P_{\mathbb{Q}}}[P(e), \mathbb{1}_e] = \mathbb{E}_{P_{\mathbb{Q}}}(P(e) \cdot \mathbb{1}_e) - \mathbb{E}_{P_{\mathbb{Q}}}(P(e)) \cdot \mathbb{E}_{P_{\mathbb{Q}}}(\mathbb{1}_e)$
- $\mathbb{E}_{P_{\mathbb{Q}}}(P(e)) = \sum_{w \in W} P_{\mathbb{Q}}(w) \cdot P_w(e)$
- $\mathbb{E}_{P_{\mathbb{Q}}}(P(e)|e) = \sum_{w \in W} P_{\mathbb{Q}}(w|e) \cdot P_w(e)$

First, we show that¹:

$$\mathbb{E}_{P_{\mathbb{Q}}}(P(e)|e) = \frac{\mathbb{E}_{P_{\mathbb{Q}}}(P(e) \cdot \mathbb{1}_e)}{P_{\mathbb{Q}}(e)} \tag{1}$$

¹Writing $P(w)$ for $P(\{w\})$, etc.

Proof.

$$\begin{aligned}
\mathbb{E}_{P_{\textcircled{a}}}(P(e)|e) &= \sum_w P_{\textcircled{a}}(w|e) \cdot P_w(e) \\
&= \sum_w \frac{P_{\textcircled{a}}(w \cap e)}{P_{\textcircled{a}}(e)} \cdot P_w(e) \\
&= \frac{1}{P_{\textcircled{a}}(e)} \sum_w P_{\textcircled{a}}(w \cap e) \cdot P_w(e) \\
&= \frac{1}{P_{\textcircled{a}}(e)} \left(\sum_{w \in e} P_{\textcircled{a}}(w) \cdot P_w(e) + \sum_{w \notin e} 0 \cdot P_w(e) \right) \\
&= \frac{1}{P_{\textcircled{a}}(e)} \left(\sum_{w \in e} P_{\textcircled{a}}(w) (P_w(e) \cdot \mathbb{1}_e(w)) \right) \\
&= \frac{\mathbb{E}_{P_{\textcircled{a}}}(P(e) \cdot \mathbb{1}_e)}{P_{\textcircled{a}}(e)}
\end{aligned}$$

□

Given (1), the rest of the proof proceeds as follows:

$$\begin{aligned}
\text{Cov}_{P_{\textcircled{a}}}[P(e), \mathbb{1}_e] > 0 &\Leftrightarrow \mathbb{E}_{P_{\textcircled{a}}}(P(e) \cdot \mathbb{1}_e) - E_{P_{\textcircled{a}}}(P(e)) \cdot \mathbb{E}_{P_{\textcircled{a}}}(\mathbb{1}_e) > 0 \\
&\Leftrightarrow \mathbb{E}_{P_{\textcircled{a}}}(P(e) \cdot \mathbb{1}_e) > E_{P_{\textcircled{a}}}(P(e)) \cdot \mathbb{E}_{P_{\textcircled{a}}}(\mathbb{1}_e) \\
&\Leftrightarrow \mathbb{E}_{P_{\textcircled{a}}}(P(e) \cdot \mathbb{1}_e) > E_{P_{\textcircled{a}}}(P(e)) \cdot P_{\textcircled{a}}(e) \\
&\Leftrightarrow \frac{\mathbb{E}_{P_{\textcircled{a}}}(P(e) \cdot \mathbb{1}_e)}{P_{\textcircled{a}}(e)} > E_{P_{\textcircled{a}}}(P(e)) \\
&\Leftrightarrow \mathbb{E}_{P_{\textcircled{a}}}(P(e)|e) > E_{P_{\textcircled{a}}}(P(e)) \quad (\text{by (1)})
\end{aligned}$$

References

- Dorst, Kevin, Levinstein, Benjamin, Salow, Bernhard, Husic, Brooke E., and Fitelson, Branden, 2021. ‘Deference Done Better’. *Philosophical Perspectives*, 35(1):99–150.
- Harsanyi, John C, 1967. ‘Games with incomplete information played by “Bayesian” players, I–III Part I. The basic model’. *Management science*, 14(3):159–182.
- Hintikka, Jaako, 1962. *Knowledge and Belief*. Cornell University Press.
- Kripke, Saul A, 1963. ‘Semantical analysis of modal logic i normal modal propositional calculi’. *Mathematical Logic Quarterly*, 9(5-6):67–96.
- Williamson, Timothy, 2008. ‘Why Epistemology Cannot be Operationalized’. In Quentin Smith, ed., *Epistemology: New Essays*, 277–300. Oxford University Press.