Base-Rate Neglect: Foundations and Implications

Dan Benjamin and Aaron Bodoh-Creed and Matthew Rabin

[Draft in Progress — Do Not Circulate]

April 15, 2018

Abstract

We extend and clarify previous formalizations of "base-rate neglect"—in which we assume that when people update beliefs from new information that they tend to downweight prior information—and explore some general implications and economic applications. We show beliefs are too moderate on average, and in fact a person may weaken her beliefs in a hypothesis even following supportive evidence. Under a natural interpretation of how it extends to dynamic settings, when an infinite flow of informative signals arrive over time, a person's beliefs will bounce around reflecting the most recent signals without converging to certainty, with a range of beliefs that is independent of the true state. Turning to economic implications, we first consider what happens when an agent is learning a "model of the world." Under mild conditions, Bayesians will learn the true model of the world, while agents subject to base-rate neglect never learn the truth and have a tendency to believe events are auto-correlated. In a persuasion setting, inducing belief updating creates a tendency towards mean reversion. Therefore, persuaders may not want to reveal even positive information when an audience has favorable current beliefs, and may share even negative information when current beliefs are unfavorable. Finally, in models where a long-run player facing a Bayesian audience is always able to build a good reputation for a long time before it eventually decays, if facing a base-rate-neglecting audience his reputation will fluctuate between good and bad in both the short run and the long run.

1 Extended Abstract

When revising beliefs in light of new information, people are prone to downweight their prior information. In this paper, we develop a formal model of such downweighting. We build the model directly from previous formulations of "base-rate neglect" (BRN), but clarify and flesh out the implications of such formulations, especially in dynamic decision-making contexts. Several of the features we identify both increase the realism of the theory relative to Bayesian updating and highlight the difficulties of applying the model. We then show some of the ways BRN differs in its implications from the Bayesian model, and how BRN might improve the realism of predictions in a number of economic settings, such as reputation building and persuasion.

In Section 2 we introduce a simple formulation of base-rate neglect. Given priors ratio $p(\theta)$ and $p(\theta')$ about two hypotheses, θ and θ', a Bayesian agent—whom we call Tommy—forms his posterior ratio after observing data D according to Bayes' Rule: $p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\Sigma_{\theta'}p(D|\theta')p(\theta')}$. An agent who suffers from base-rate neglect—whom we call Saki—uses likelihood information exactly as Tommy does, but underweights her prior beliefs: $p_\alpha(\theta|D) = \frac{p(D|\theta)p(\theta)^\alpha}{\Sigma_{\theta'}p(D|\theta')p(\theta')^\alpha}$, where $\alpha \in [0,1)$. If $\alpha = 0$, Saki completely ignores base rates when she updates. When $\alpha \in (0,1)$, Saki under-uses, but does not completely neglect, her priors. This straightforward and intuitive formulation has been previously estimated in empirical tests of base-rate neglect.

In Section 2 we also discuss some of the evidence for base-rate neglect, and in the process explore the meaning of, the psychology of, and several problems with this formalization. In addition to the more commonly noted implications of base-rate neglect, we present evidence of some of its surprising features. Along with the better-known implication of BRN that it can lead to violations of the conjunction principle—that a person should always place at least as high a probability on a set of possibilities than on any subset of those possibilities—there is a striking prediction of the above formula that is virtually unmentioned in presentations of BRN: if Saki's prior belief strongly favors one hypothesis and she observes weak evidence *in favor* of that hypothesis, then her posterior belief may favor that hypothesis *less* strongly. As one of the rare papers that includes treatments where base rates and new information both lean in the same direction, Griffin and Tversky (1992) in fact provide evidence for this extreme moderation effect.

Section 2 also explores some other challenging features of the model. As with other non-Bayesian models, the effects of base-rate neglect inherently depend on the framing of hypotheses. Suppose that Saki is a manager assessing two of her employees, Heidi and Tarso. If a manager is updating her beliefs about the two possibilities whether Heidi is more effective than Tarso versus not more effective, she will reach different conclusions about whether Heidi is effective than if she is updating about three possibilities: Heidi is more effective, less effective but competent, or incompetent. We formalize how beliefs about the likelihood of these different potential conclusions always exhibit subadditivity, which means an event is viewed as more likely if it is framed in terms of disjoint sub-events. Indeed, base-rate neglect can also lead to violations of conjunction across these two scenarios: following information hinting that Tarso is more competent than Heidi, Saki might put more weight on Heidi being competent but inferior in the second scenario than she would put weight on Heidi being competent in the first scenario. It is of course logically impossible for the probability of Heidi being competent to be lower than her being competent but inferior to Tarso. We show how base-rate neglect can cause such violations whenever a signal is relatively unlikely in event E, but is very likely in a low probability subevent of E. Conjunction violations happen because Saki neglects the low probability of the subevent, but this is of course a logical contradiction. A related problematic feature of base-rate neglect is that Saki will modify her beliefs based on information that turns out not to be diagnostic at all. For example, when Saki is trying to assess whether an employee is competent and observes data she anticipates to be diagnostic but turns out not to be, she is apt to change her beliefs. We discuss how one might go about making judgments about when Saki updates her beliefs.

Section 2 also explores some ways in which BRN as it has been previously formulated is manifestly incomplete: It says what Saki thinks *retrospectively* after she sees evidence, but does not specify what she thinks *prospectively* her beliefs will be when she gets additional information. Retrospective and prospective beliefs are, by construction, equivalent for a Bayesian. But as discussed and modeled in Benjamin, Rabin, and Raymond (2015) in the context of a different error, there is little reason to believe that the two will be consistent for cognitive errors.[1] Although prospective beliefs are not much implicated in the applications discussed in this paper, completeness and the potential for further applications demands an assumption. Without much empirically to go on, we complete the model by assuming that Saki thinks that her future updating will be Bayesian. This in turn means that Saki believes her beliefs will obey the Law of Iterated Expectations: if she currently believes there is an 80% chance of something, she'll anticipate 80% beliefs tomorrow, not the more moderate beliefs she actually will believe.

In Section 3, we study a stylized dynamic environment where an agent observes a sequence of signals that are informative about the underlying hypothesis. We assume that each time Saki receives an informative signal, she updates her beliefs to form a posterior. The next time she receives a signal, she (under-)uses this updated posterior as her prior belief when she next updates. To show some of the long-run implications of this dynamic, we explore some results in the case where Saki receives an infinite stream of i.i.d. signals. The net effect is that Saki increasingly neglects past information as new signals arrive. Indeed, because the influence of a signal on Saki's current belief is exponentially declining in the number of intervening signals, Saki's belief after any number of signals is always determined by the informational equivalent of a finite number of signals, with both her initial priors and early signals eventually having no influence. Whereas Tommy eventually learns the correct hypothesis with virtual certainty in this setting, Saki will never become confident in the correct hypothesis, and her beliefs will continue to oscillate. Although the long-run *frequency* of different beliefs depend on the truth, the *range* of beliefs she returns to does not: Saki will occasionally return to believing in each hypothesis strongly when she happens to get a string of signals supporting that belief.

In Section 4, we explore some of the implications of our formulation of base-rate neglect for some existing contexts and notions of belief updating from sequences of signals. In particular, we consider how Saki's model of the world evolves, where by "model of the world" we mean her predictions of future random events. Since the model of the world is the hypothesis she is testing and she never becomes confident in the correct hypothesis, Saki's beliefs about the model of the

[1] Besides Benjamin, Rabin, and Raymond (2015), the only paper we are familiar with that explores the relationship between prospective and retrospective beliefs is He and Xiao (mimeo). They explore the implications of imposing consistency between the two, and show how some features of well-known biases can (and, mostly, cannot) be captured in such consistent models.

world most likely to be correct continue to evolve as she gathers successive signals. Moreover, her beliefs about the most likely model of the world tend to reflect her most recent observations (and neglect past observations), which can cause her model of the world to change quickly and to take extreme forms.

For example, suppose Saki is determing which side of a bet to take on a future basketball game. Because BRN causes a recency bias, Saki will believe that after observing a team win a game that the team is more likely to win its next game and vice versa. This is not because Saki believes in autocorrelated performance across games, referred to as a "hot hand" by Camerer (1989). Instead, it is because Saki's beliefs about the teams's permanent ability are much more optimistics after observing a win than after observing a loss. More generally, in a variety of settings Saki will conclude from recent success or failure that future success is likely. Moreover, and unlike in a model of autocorrelated performance, Saki believes that recent success predicts a high likelihood of success into the indefinite future.

In a canonical normal-normal updating regime, we also find that BRN causes Saki's beliefs to behave as if she had adaptive expectations. Agents with adaptive expectated equate the expected value of a random variable with a discounted sum of prior realizations of that variable. While adaptive expectation models has often been found to be a useful description of how the beliefs of individuals evolve (for a review, see Fuster et al. (2010)), the classical models could not say much about how individuals respond to endogenous structural changes in the economy since these models typically violated the rational expectations hypothesis. Base-rate neglect can predict that Saki's beliefs exhibits a pattern similar to that of adaptive expectations when the economy's structure is stable, but because Saki is assumed to have a correct model of how the economy works, she is responsive to endogenous changes in the economy.

We explore some economic implications of BRN in cases where a rational economic actor interacts with Saki in Sections 5 and 6. We show how each of the properties of beliefs discussed above—the non-convergence, the perpetual fluctuations, and the independence of the support from the truth—has economic implications. In Section 5, we consider a fully rational "persuader" attempting to persuade an "audience." The persuader can influence the audience's beliefs by choosing whether or not to reveal a signal. A revealed signal is verifiable, but the existence, absence, and nature of an unrevealed signal is not verifiable. If the audiences are Bayesian, the unique sequential equilibrium is for information to be revealed if and only if it is good. In equilibrium, the audience deduces that the absence of a revealed signal implies either a bad signal or no information. We analyze the contrasting implications of BRN. We assume that, despite neglecting base rates, Saki is strategically sophisticated: she understands how the motives and available actions of would-be persuaders may influence what messages they wish to send—or whether they wish to send information at all. We show that—despite this sophistication about what silence may mean—that a persuader who is happy with status-quo beliefs may prefer to not reveal even favorable information to Saki so as to prevent the negative updating that can occur even for a good signal. When this is the case, we show that despite strategic sophistication, the moderation effect means that the classical "unwinding" result for verifiable-information revelation does not hold. On the flipside, if Saki's prior is unfavorable to the persuader, then the persuader might be willing to reveal even bad signals, because (again due to the moderation effect) the persuader can muddy Saki's beliefs by communicating.

In Section 6 we examine the implications of BRN for reputation-building. We consider the prototypical setting first studied by Fudenberg and Levine (1992): a long-run, patient firm can be either a "strategic" player that decides each period whether to "shirk" or "work" on unobservably high quality that period, or with small probability it can be a "committed type" that automatically works each period. Each period, a new consumer decides whether or not to buy without knowing current quality, but after observing the history of product quality from previous periods, which might be informative about whether the firm is a strategic player or a committed type. Fudenberg and Levine (1992) show that, if the consumers are Tommys, then the (strategic) firm will mimic the committed type (or do as well as if it mimicked the committed type) for a very long time. Cripps, Mailath, and Samuelson (2004) in turn show that in equilibrium a strategic firm's reputation is impermanent: the firm will work with high probability initially, but as time goes strategic firms are found out and begin to shirk consistently. We show that if the consumers are Sakis, then in equilibrium consumers will never become extremely confident about the firm's type and the firm's reputation will never be completely and permanently destroyed. In this context of unobservable effort with noisy results, neither the initial (long duration of) good reputation nor the eventual decay in that reputation necessarily holds.

In Section 7 we discuss the relationship between BRN and other types of biases. While there can be interesting interactions between BRN and biases such as the Law of Small Numbers (LSN) (Rabin (2002)) and Nonbelief in the Law of Large Numbers (NBLLN) (Benjamin, Rabin, and Raymond (2015)), the most necessary comparison is with models of confirmatory bias. As modeled in Rabin and Schrag (1999), confirmatory bias formalizes the psychology whereby people tend to misread evidence as supporting their currently held beliefs. "Currently held" is interpreted in their model as the person's beliefs about which of two hypotheses is more likely. As such, confirmatory bias leads to people being *over*-influenced by priors entering a period.