Updating: Learning versus supposing

Jiaying Zhao (jiayingz@princeton.edu)
Department of Psychology, Green Hall,
Princeton University, NJ 08540 USA

Vincenzo Crupi (vincenzo.crupi@unito.it)
Department of Philosophy, University of Turin, Turin, Italy

Katya Tentori (katya.tentori@unitn.it)
DiSCoF, CIMeC, University of Trento, Rovereto, Italy

Branden Fitelson (branden@fitelson.org)
Department of Philosophy, Rutgers University,
New Brunswick, NJ 08901 USA

Daniel Osherson (osherson@princeton.edu)
Department of Psychology, Green Hall,
Princeton University, NJ 08540 USA

Abstract

Bayesian orthodoxy posits a tight relationship between conditional probability and updating. Namely, the probability of an event *A* after learning an event *B* should equal the conditional probability of *A* given *B* prior to learning *B*. We examine whether ordinary judgment conforms to the orthodox view. In three experiments we found substantial differences between the conditional probability of an event *A* supposing an event *B* compared to the probability of *A* after having learned *B*. Specifically, supposing *B* appears to have less impact on the credibility of *A* than learning that *B* is true. Thus, Bayesian updating seems not to describe the relation between the probability distribution that arises from learning an event *B* compared to merely supposing it.

Keywords: belief updating, reasoning, probability

Introduction

Let Pr_1 represent the beliefs of an idealized agent who is considering at time 1 the credibilities of events over an outcome space Ω (finite, for simplicity). Suppose that for some event $B \subseteq \Omega$ with $Pr_1(B) > 0$ experience intervenes at time 2 to convince the agent that *B* is (definitely) true. What new distribution Pr_2 should embody the agent's revised beliefs? The Bayesian response (Hacking, 2001, Ch. 15) is that Pr_2 should be the result of conditioning[1] Pr_1 on *B*, that is:

(1) BAYESIAN UPDATING:

If $B \subseteq \Omega$ is learned between times 1 and 2 (and nothing else relevant is learned) then for all events $A \subseteq \Omega$, $Pr_2(A) = Pr_1(A \mid B)$ (provided that $Pr_1(B) > 0$).

It is easy to check that Pr_2 as defined by (1) is a genuine probability distribution and that $Pr_2(B) = 1$ (as expected). Also, (1) is a consequence of compelling axioms on belief change (Gardenfors, 1988, §5.2), and its violation exposes the agent to sure-loss betting contracts (Harman, 1999, §4.12).

Such normative virtues suggest a psychological question. One way of formulating (1) is that *supposing* an event *B* should have the same impact on the credibility of an event *A* as *learning B*. Is this true for typical assessments of chance? For example, is the judged probability of a Democratic victory in 2012 *supposing* that Hilary Clinton is the vice presidential candidate the same as the judged probability of a Democratic victory in 2012 after *learning* that Clinton, as a matter of fact, is the vice presidential candidate?

The foregoing question is orthogonal to the provenance of conditional probability in the mind, that is, to the way such probabilities are mentally computed. Thus, even if people fail to respect the standard definition:

$$Pr(A \mid B) \overset{\text{def}}{=} \frac{Pr(A \cap B)}{Pr(B)}$$

it is still possible for (1) to hold.[2] All that matters is whether the same degree of confidence in event *A* is reached when supposing event *B* compared to learning it. On the other hand, recent literature on conditional reasoning has suggested a difference between supposing vs. learning the antecedent of a conditional (Oaksford & Chater, 2007). We investigated the difference between learning and supposing in three experiments.

Experiment 1

Participants

Forty undergraduates (27 female, mean age 19.4 yrs, SD = 1.3) from Princeton University participated in exchange for course credit.

[1] By conditioning we refer to simple or strict Bayesian updating, rather than associative learning.

[2] An earlier study focussed on the fidelity of the standard definition (above) to the numbers people report as conditional probabilities (Zhao, Shah, & Osherson, 2009).

Materials

Five decks of cards served as stimuli, each with 20 cards. Each card presented an animal and a colored square on one side and was blank on the other side. The animal was marked on the bottom half of the card and could be either a dog or a duck. The colored square was marked on the top half and could be either green or yellow. Thus, each deck contained four types of cards: green dog, green duck, yellow dog, and yellow duck. Table 1 summarizes the respective frequencies of the types of cards for each deck.

Table 1: Number of cards in each deck used in Experiment 1.

Deck	Green Dog	Green Duck	Yellow Dog	Yellow Duck
1	5	4	6	5
2	9	2	6	3
3	4	8	2	6
4	7	8	2	3
5	3	3	6	8

Procedure

There were two conditions in the experiment: *learn* and *suppose*, each with 20 participants. In both conditions, the five decks were presented to the participant in random order. For each deck, the experimenter first showed the cards to the participant, with the animals and colors in plain view. Cards were presented briefly (around 0.5 second apiece) to prevent counting. After all cards in the deck were presented, the participant shuffled the deck, drew one card at random, and put it on the table blank side up. Thus, neither the participant nor the experimenter knew what the card was.

The procedure then differed between the two conditions. In the learn condition, the experimenter covered the card drawn from the deck, turned the card over while still covered, and then revealed one half of the card to the participant. Whether the animal or the color was thereby revealed was randomly determined. If the revealed half was an animal then the participant estimated the probability that the unrevealed half was a certain color; whether they were asked for the probability of "green" versus "yellow" was determined randomly. If the revealed half was a color then the participant estimated the probability that the unrevealed half was a certain animal; whether they were asked for the probability of "dog" versus "duck" was determined randomly. The covered half was never revealed to the participant. This procedure was repeated for all five decks.

The suppose procedure was identical to the foregoing up to placing one card from the shuffled deck face down on the table. In the suppose procedure, neither side of the card was revealed, and the experimenter proceeded instead to ask a question of the form: "What is the probability that so-and-so appears on the card supposing that such-and-such appears?" The content of the question (so-and-so and such-and-such) was determined by yoking each suppose participant to the immediately preceding participant, who was in the learn condition. Specifically, if for decks 1 through 5 the learn participant was asked for the probabilities of $A_1 \ldots A_5$ upon learning $B_1 \ldots B_5$ then the suppose participant was asked for the conditional probabilities $Pr(A_1 \mid B_1) \ldots Pr(A_5 \mid B_5)$ in the order corresponding to the presentation of the five decks to the learn participant.

Thus, the first participant was assigned to the learn condition, the second to the suppose condition (and yoked to the first participant), and likewise for succeeding pairs of participants. The crucial difference was that participants in the learn condition estimated A after learning B, whereas participants in the suppose condition estimated A while supposing B.

Results and Discussion

We computed three statistics over the five probabilities that a given participant produced, namely, (a) the average of the five raw responses, (b) the average absolute deviation from 0.5, and (c) the average absolute deviation of a response from the objective probability of the event under consideration (where the objective probability was derived from the composition of the deck employed in that trial). Statistic (b) quantifies confidence inasmuch as extreme probabilities signify presumed knowledge about an event whereas 0.5 represents ignorance. The statistics produced by the two groups were then compared via paired t-tests. There were thus 20 pairs, defined by yoking each *suppose* participant to his/her *learn* participant.

As seen in row (a) of Table 2, the average responses across the 20 learn-suppose pairs were virtually identical [$t(19) = 0.60$, $p = 0.56$, $d = 0.13$]. Row (b) shows, however, that the absolute deviation from 0.5 was reliably greater for the suppose group compared to learn group [$t(19) = 2.61$, $p = 0.02$, $d = 0.58$]. The absolute deviation from objective probability also differed reliably between the learn and suppose conditions [$t(19) = 4.15$, $p < 0.001$, $d = 0.93$] with more accurate responses from the learn participants; see row (c). Moreover, in 16 of the 20 pairs, learn participants were more accurate than suppose participants ($p = 0.01$ by binomial test).

Table 2: Comparison of learn and suppose groups in Experiment 1

Statistic	Learn	Suppose	p
(a) Raw estimate of $Pr(A\|B)$	0.50(0.10)	0.49(0.12)	0.56
(b) Abs. dev. from 0.5	0.14(0.05)	0.19(0.06)	0.02
(c) Abs. dev. from $Pr(A\|B)$	0.09(0.05)	0.16(0.08)	0.00

Means for the two groups, relative to various statistics. Standard deviations are given in parentheses; p-values reflect paired t-tests ($N = 20$). (Abs. dev. = Absolute deviation)

The results of Experiment 1 thus suggest limitations to the Bayesian model of updating, at the descriptive level. To check the robustness of our findings, we repeated Experiment 1 with new decks, involving different frequencies of the four events.

Experiment 2

Participants

A new group of forty undergraduates (26 female, mean age 19.5 yrs, SD = 1.8) from Princeton University participated in exchange for course credit.

Materials and procedure

The procedure was identical to Experiment 1 except for the use of five different decks shown in Table 3.

Table 3: Number of cards in each deck used in Experiment 2.

Deck	Green Dog	Green Duck	Yellow Dog	Yellow Duck
1	9	2	1	8
2	2	8	9	1
3	7	1	3	9
4	3	8	7	2
5	8	3	2	7

Results and Discussion

We computed the same statistics as in Experiment 1; see Table 4. As before, there was virtually no difference in the average responses of the suppose versus learn groups. And once again, suppose participants were less accurate than learn participants in terms of absolute deviation from the objective value; 16 of the twenty learn/suppose pairs showed this pattern ($p = 0.01$ by binomial test). This time, however, learn rather than suppose participants issued more extreme probabilities; see row (b) of Table 4.

Table 4: Comparison of learn and suppose groups in Experiment 2

Statistic	Learn	Suppose	p
(a) Raw estimate of $Pr(A\|B)$	0.50(0.14)	0.48(0.10)	0.53
(b) Abs. dev. from 0.5	0.23(0.05)	0.18(0.06)	0.00
(c) Abs. dev. from $Pr(A\|B)$	0.13(0.04)	0.21(0.08)	0.00

Means for the two groups, relative to various statistics. Standard deviations are given in parentheses; p-values reflect paired t-tests ($N = 20$). (Abs. dev. = Absolute deviation)

Why did suppose participants issue more extreme probabilities than learn participants in Experiment 1 while the reverse is true in Experiment 2? Tables 1 and 3 report the objective distributions of the cards in the two experiments, and reveal greater extremeness for the second experiment compared to the first. So, the switch in extremeness might be a corollary to the greater accuracy of the learn compared to suppose groups.

In sum, the results of Experiment 2 reveal once again a gap between learning and supposing that is not foreseen by Bayesian updating.

Experiment 3

In Experiments 1 and 2, probabilities were grounded in frequencies and therefore *extensional*. The third experiment was designed to evaluate the impact of learning versus supposing in an *intensional* setting involving probabilities of non-repeatable events. In particular, participants in the third experiment specified their confidence (as a probability) that Bill Clinton won/lost a specified state given that he won/lost another state in the 1992 presidential election.

Participants

A new group of sixty undergraduates (41 female, mean age 20.4 yrs, SD = 1.9) from Princeton University participated in exchange for course credit.

Materials

A deck of 50 cards served as stimuli. One side of a given card was marked with a U.S. state, the other side left blank.

Procedure

As in the previous experiments, there was a learn and a suppose condition, each with 20 participants. In both conditions, the participant examined the deck then shuffled it and placed two cards face down on the table (without looking at them). Despite the appearance of randomness, the experimenter examined but then ignored the contents of the cards, and instead asked about two states from a pre-selected list. The list consisted of 20 swing states (electoral outcome not easily predictable); the two swing states figuring in a given trial were drawn randomly from the list.[3]

In the learn condition the experimenter picked up one of the two drawn cards and looked at its underside (preventing the participant from seeing the content). The state was announced (actually, the announced state was preselected from the list of 20 swing states), and then the electoral outcome for that state was determined by consulting a website. Specifically, with the participant watching, the experimenter discovered the outcome for that state via `http://uselectionatlas.org/RESULTS/`, and showed the result to the participant. Note that the participant was only shown whether Clinton won or lost the specified state; information about other states was masked. The experimenter then examined the underside of the remaining card, announced this second state (actually, preselected from the list of swing states), and asked the participant to estimate the probability of Clinton winning or losing that state. The framing of the question in terms of winning or losing was consistent with the outcome for the first state. For example, if Clinton won the first state then the participant estimated the probability of Clinton winning the second state, and likewise for losing. The two cards were then put aside, never revealed to the participant. This procedure was performed five times per participant.

In the suppose condition, each participant was yoked to the immediately preceding learn participant. To start the trial, the experimenter announced that it was a winning (or losing) round, meaning that the participants were to estimate the

[3] The swing states were taken to be AL, AZ, GA, ID, IN, KS, KY, LA, MI, MN, MO, MS, MT, NC, ND, NM, OH, TN, VA, WV.

probability that Clinton won (or lost) the second state, supposing that he won (or lost) the first state. The choice of framing (win/lose) appeared to be random, but in fact matched the questions in the learn condition. To finish the trial, the participant shuffled the deck and placed two cards face down on the table (without looking). The experimenter pretended to look at the undersides of the two cards and asked the participant to estimate the probability of Clinton winning (or losing) the second state supposing that he won (or lost) the first state. The two states were yoked to those in the learn condition. The two cards were then set aside, never revealed to the participant. This procedure was performed five times (yoked to the preceding learn participant).

A third group ($N = 20$) served as a control condition in which just $Pr(A)$ was estimated (no conditioning event B was evoked). In this condition, each participant was yoked to the preceding suppose and learn participants, and gave probabilities to the five states that were target events A. For each trial, the experimenter announced that it was a winning (or losing) round, meaning that the probability to be estimated was that Clinton won (or lost) the state in question. The framing was yoked to the questions in the suppose and learn conditions. The participant then shuffled the deck and placed one card face down on the table. The experimenter pretended to look at the card and asked the participant to estimate the probability of Clinton winning (or losing) the state. The procedure was performed for each of the five states.

Results and Discussion

As seen in row (a) of Table 5, the average responses of the learn participants were reliably higher than those of the suppose participants [$t(19) = 4.41$, $p < 0.001$, $d = 0.99$]. Row (b) shows that the learn group offered more extreme probabilities than the suppose group [$t(19) = 3.11$, $p < 0.01$, $d = 0.69$].

Table 5: Comparison of learn and suppose groups in Experiment 3

Statistic	Learn	Suppose	p
(a) Raw estimate of $Pr(A\|B)$	0.64(0.09)	0.53(0.10)	0.00
(b) Abs. dev. from 0.5	0.21(0.06)	0.15(0.06)	0.00
(c) Quadratic penalty	0.18(0.08)	0.25(0.08)	0.02

Means for the two groups are presented, relative to various statistics. Standard deviations are given in parentheses; p-values reflect paired t-tests ($N = 20$). (Abs. dev. = Absolute deviation)

To quantify the accuracy of the probability assigned to event A upon learning or supposing B, we computed the *quadratic penalty* for $Pr(A)$. To illustrate, the quadratic penalty for $Pr(\text{wins Virginia})$ is $(1 - Pr(\text{wins Virginia}))^2$ in the event that Clinton won Virginia, and it is $(0 - Pr(\text{wins Virginia}))^2$ in case he lost. (Note that the conditioning event B was true in every case.) Thus, low penalty signifies accuracy of a stochastic forecast whereas high penalty signifies inaccuracy; assigning the noncommittal probability 0.5 guarantees a penalty of 0.25, indicating ignorance. The quadratic penalty was introduced by Brier (1950); see Predd et al. (2009) for a justification of its use in measuring accuracy. For every participant, we computed her average quadratic penalty over the five trials. Row (c) of Table 5 shows that learners were closer to the truth than supposers were [$t(19) = 2.57$, $p = 0.02$, $d = 0.58$]. This holds for 17 of the 20 pairs of participants ($p = 0.01$ by binomial test). It is striking that the mean quadratic penalty for the suppose group is almost exactly 0.25, the accuracy level guaranteed by issuing 0.5 probabilities.

We note that a majority (60%) of the pairs figuring in the experiment had consistent outcomes in the election (Clinton winning both or losing both). For the learn condition, the average probabilities assigned to consistent and inconsistent pairs were 0.72 and 0.50, respectively, whereas they were 0.55 and 0.48 for the suppose condition. A two-way ANOVA reveals a reliable interaction, the difference between the learn probabilities exceeding that for the suppose probabilities [$F(1,19) = 17.7$, $p < .001$]. Since a majority of pairs were consistent (as noted above), these facts explain the lower quadratic penalty for learn participants, and highlight their greater sensitivity to the conditioning event B.

Finally, in the control condition, the average raw estimate of $Pr(A)$ was 0.51 ($SD = 0.13$). This is close to the 0.53 estimate of $Pr(A \mid B)$ in the suppose group [$t(19) = 0.51$, $p = 0.61$, $d = 0.12$] but reliably different from the 0.64 estimate of the learn group [$t(19) = 4.09$, $p < 0.001$, $d = 0.91$]. The quadratic penalty for the control condition was 0.29 ($SD = 0.09$), reliably different from learn [$t(19) = 4.18$, $p < .001$, $d = 0.94$] but not suppose [$t(19) = 1.50$, $p = 0.15$, $d = 0.34$]. These results indicate once again that the conditioning event B had greater impact on the judgements of the learn participants compared to suppose.

General Discussion

Bayesian updating (1) seems not to describe the relation between the probability distribution that arises from learning an event B compared to merely supposing it. For, in our three experiments, the probabilities that issue from learning B are more accurate than those resulting from conditioning, and they also differ in their deviation from 0.5. In Experiment 3, moreover, the average probabilities in the two groups differed significantly.

In the latter experiment, learn participants seem to have made greater use of the conditioning event B than did suppose participants. This is revealed by the greater difference in updated compared to prior probabilities for A in the learn compared to the suppose conditions. Specifically, learn estimates were reliably higher than the prior, suggesting that learn participants interpreted a win [loss] of one swing state to increase the chance of a win [loss] of another. In contrast, suppose participants' estimates of $Pr(A \mid B)$ were almost identical to the control group's $Pr(A)$.

Insensitivity to B may reflect a deficit of imagination, the suppose participants being unable to simulate the effect of genuinely believing *B*. In fact, Bayesian updating imposes a heavy burden on the reasoner's ability to foresee the impact of experience. Suppose that lions are discovered roaming your neighborhood; can you anticipate the probabilities you would attach to other events if such startling circumstances actually came to pass? Analogous difficulties arise when attempting to predict future affective states (Wilson & Gilbert, 2003).

At the normative level, the Bayesian doctrine (1) is supported by the considerations mentioned in the introduction, yet it remains contentious (see, e.g., Bacchus, Kyburg, and Thalos (1990)). Recent work has begun to provide a necessary critique of Bayesian inference (Jones & Love, 2011). The Bayesian doctrine may also prove to be unsuited to situations in which the agent, albeit rational, loses track of her position in time or space (Arntzenius, 2003). But the debate about (1) might be of limited relevance to the typical transition from one probability distribution to another. Such transitions need not depend on adding an event *B* to one's beliefs without probabilistic qualification. Rather, experience might lead us to revise our confidence in *B* without driving it to zero or one. The rule proposed by Jeffrey (1983) is suited to this kind of case. Recent work has begun to examine Jeffrey's rule from the psychological point of view (Over & Hadjichristidis, 2009; Zhao & Osherson, 2010).

Acknowledgments

Osherson acknowledges support from the Henry Luce Foundation.

References

Arntzenius, F. (2003). Some problems for conditionalization and reflection. *The Journal of Philosophy*, 356-370.

Bacchus, F., Kyburg, H., & Thalos, M. (1990). Against conditionalization. *Synthese*, *85*, 475–506.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1-3.

Gardenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. Cambridge MA: MIT Press.

Hacking, I. (2001). *An Introduction to Probability and Inductive Logic*. Cambridge UK: Cambridge University Press.

Harman, G. (1999). *Reasoning, meaning and mind*. Oxford UK: Oxford University Press.

Jeffrey, R. C. (1983). *The Logic of Decision (2nd Edition)*. Chicago IL: The University of Chicago Press.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, *34*, 169-231.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.

Over, D., & Hadjichristidis, C. (2009). Uncertain premises and jeffrey's rule. *Behavioral and Brain Sciences*, *32*, 97 - 98.

Predd, J., Seiringer, R., Lieb, E. H., Osherson, D., Poor, V., & Kulkarni, S. (2009). Probabilistic coherence and proper scoring rules. *IEEE Transactions on Information Theory*, *55*(10), 4786 - 4792.

Wilson, T., & Gilbert, D. (2003). Advances in experimental social psychology. In M. P. Zanna (Ed.), (Vol. 35, p. 345-411). Academic Press.

Zhao, J., & Osherson, D. (2010). Updating beliefs in light of uncertain evidence: Descriptive assessment of Jeffrey's rule. *Thinking & Reasoning*, *16*, 288-307.

Zhao, J., Shah, A., & Osherson, D. (2009). On the provenance of judgments of conditional probability. *Cognition*, *113*(1), 26 - 36.