Notes and Comment

Commentary on Wolford, Taylor, and Beck: The conjunction fallacy?

MAYA BAR-HILLEL

The Hebrew University, Jerusalem, Israel

Although $P(A \& B | X)$ can never exceed $P(A | X)$ (the conjunction rule), it is possible for $P(X | A \& B)$ to exceed $P(X | A)$. Hence, people who rank $A \& B$ as more probable than A are not necessarily violating any normative rule if the ranking is done in terms of the probability of these events to yield an event X . Wolford, Taylor, and Beck (1990) have argued that this indeed is what happens in some problems (e.g., Tversky & Kahneman's [1983] Linda problem). The claim made here is that the Linda problem is hard to reconcile with this interpretation; that there is little if any evidence that subjects utilize this interpretation; and that in any case, representativeness can account for all Linda problem results.

A conditional probability, $P(A | B)$, and its inverse, $P(B | A)$, are often—though not invariably—easy to confuse. For example, $P(\text{symptom} | \text{malady})$ is often confused with $P(\text{malady} | \text{symptom})$ (Eddy, 1982), but it is hard to confuse $P(\text{male} | \text{senator})$, which is close to one, with $P(\text{senator} | \text{male})$, which is close to zero (Bar-Hillel, 1984). In a manner of speaking, all probabilities are conditional—conditional on the sample space in question, on the so-called “universal event.” But when the conditioning event is held constant, it becomes implicit, in mind as well as in notation. Hence, the laws governing the probabilities of compound events which are conditioned on a constant sample space are simply the familiar laws of “unconditional” probabilities. On the other hand, if it is the conditioned event that is held constant in some context, while the conditioning events are changed, then the laws governing the corresponding probabilities are quite different.

Consider an example. The extension rule states that if A is contained in B then $P(A | X) \leq P(B | X)$.¹ This rule has no analogue for the inverse probabilities. Indeed, if A is contained in B , no constraints whatsoever are set on the order between $P(X | A)$ and $P(X | B)$ —either one can be larger than the other, or they can be equal. A graphical example appears in the accompanying Venn diagram (Figure 1). A special form of the extension rule is the conjunction rule, which states that $P(A \& B | X) \leq P(A | X)$. No analogous rule constrains the order between $P(X | A \& B)$ and $P(X | A)$. For a numerical example, see Wolford, Taylor, and Beck (1990).

In a study that became an instant classic, Tversky and Kahneman (1983) showed that under certain conditions, people produce probability judgments that predictably and

systematically violate the conjunction rule. In particular, when people are given a description of an individual L (called, say, Linda), who sounds much like an F (say, feminist), and hardly at all like a B (say, bank teller), they tend to judge Linda more likely to be an $F \& B$ (i.e., a feminist bank teller) than a B (i.e., a bank teller). Recently, Wolford et al. (1990) addressed themselves to this result, which Tversky and Kahneman (1983) called the conjunction fallacy. In defense of some subjects alleged to have committed this fallacy in the Linda problem and its likes, Wolford et al. (1990) suggested that the subjects were judging $P(L | F \& B)$ versus $P(L | B)$, not $P(F \& B | L)$ versus $P(B | L)$, as Tversky and Kahneman's (1983) analysis assumes. To the extent that this might have been the case, the subjects would have violated no formal rule.

Wolford et al.'s (1990) defense is twofold. First, they argue that their alternative interpretation is (normatively) appropriate for the Linda problem; second, they attempt to show that it is descriptively appropriate as well—namely, that it is indeed the interpretation favored by the majority of subjects receiving this problem.

Their normative argument is couched in terms of “unknown versus known outcomes.” The “model for unknown outcomes . . . applies to events in which we are trying to predict an unknown outcome” (Wolford et al., 1990, p. 49), and is simply the standard probability model. On the other hand, the “model for known outcomes . . . [applies when] we know that an outcome has occurred and our job is to discover the correct outcome” (p. 49). The model for known outcomes is presented via the following example:

A robbery . . . occurred in a small town and the local police chief knows all of the possible suspects. She first learns that dynamite was used in the robbery. Person X is unlikely to use dynamite, so the police chief does not place X very high on the suspect list. The police chief then discovers that the robber had bright red hair. Since X has bright red hair, he moves to a much higher spot on the suspect list. (p. 49)

Figure 1. A Venn diagram in which $P(X | A \& B) = 1$, while $P(X | A), P(X | B) < 1$.

Correspondence should be addressed to Maya Bar-Hillel, Department of Psychology, The Hebrew University, Jerusalem 91905, Israel.

Wolford et al. correctly add that this is "an appropriate way to revise the probability of X 's having committed the crime." This, of course, is because we are dealing with $P(X|E_1)$ versus $P(X|E_1 \& E_2)$ —where E_1 represents the robber's having used dynamite, and E_2 represents the robber's having had red hair—and the latter can coherently be greater than the former. Note, however, that this possibility in no way depends on anything's being "known." The police chief need not know all the possible suspects at all, and X need not even be one of them. Even if the police chief has no suspects yet, and X is merely her next-door neighbor, the analysis above is perfectly valid. In the same vein, if the Linda problem had included an explicit statement to the effect that Linda was known to be doing something on the list of possible outcomes, its normative analysis would not have changed at all.

It appears that though Wolford et al. (1990) saw the crucial difference between their two models as consisting of whether or not the outcomes were known, the models in fact differ simply in terms of whether the target event (the one that is the focus of the judgmental task) is the conditioned one or the conditioning one—namely, in terms of whether one is judging $P(X|E)$, or $P(E|X)$. We shall now see which model is more appropriate for the Linda problem.

Consider first Wolford et al.'s (1990) analysis of the Linda problem.

The Linda scenario seems neutral with respect to the models outlined above. Subjects are not told that Linda's fate is known or that one of the alternatives represents Linda. Subjects are not told the contrary, either. It seems to us that the most appropriate framing would be to assume that the outcome is unknown and that the unknown model applies, but subjects easily could assume that the three alternatives [feminist, feminist bank teller, and bank teller] represent three different individuals and that one of the alternatives represents Linda. In the latter case, rating [feminist bank teller] higher than bank teller alone does not represent a fallacy. (p. 50)

To check which interpretation subjects were actually using, Wolford et al. (1990) asked "a group of 12 psychology graduate students and faculty members to rate the ... problems as to whether the outcomes were known or unknown." From the fact that the outcomes were rated closer to "known" than to "unknown" (an average of 2.4 on a 7-point scale ranging from "already known and might be ... listed" to "would not be known until sometime in the future"), Wolford et al. concluded that "for most ... subjects, the known model would be appropriate, and rating the conjunction higher than one of the constituents would not necessarily be a fallacy" (pp. 52–53).

Translated into the terms of conditional probabilities, Wolford et al. (1990) are apparently saying that most subjects presented with the original Linda problem could well have been judging $P(L|F)$, $P(L|F \& B)$, and $P(L|B)$ —not $P(F|L)$, $P(F \& B|L)$, and $P(B|L)$. Moreover, they seem to be saying that both judgments would have been legitimate, since subjects were simply not told whether the outcomes were known or not (a consideration whose irrelevance I have argued above).

Since the legitimacy of employing the known model cannot be derived from the judgment that the outcomes are known, let us consider its legitimacy directly. The question is: Do subjects see their task as that of matching occupations to candidates ("What is Linda more likely to be: a feminist bank teller or a bank teller?") or of matching candidates to occupations ("Who is more likely to be the feminist bank teller: Linda or Carol?")? It is certainly normatively possible for Linda to be more likely to be a feminist bank teller (than Carol), while simultaneously being less likely to be a bank teller (than Carol). Thus, "feminist bank teller" can be ranked higher than "bank teller" in terms of "more likely to be Linda (than Carol)." But it is hard, if not impossible, to see how subjects given the original Linda problem—which explicitly asks them to match occupations to a person, and which mentions only one person, Linda—can come to think that they are matching persons to occupations.

What, then, have the authors shown?—first, that ranking the probability of $F \& B$ as higher than that of F is not necessarily an error, in cases in which the ranking is made in terms of these alternatives' probability of yielding L (i.e., $P(L \bullet)$); this is unarguably correct. Second, that in the Linda problem, "subjects easily *could* assume" (*italics mine*) that their task is to rank the alternatives in just these terms. The authors have actually not shown this; they have merely stated it (see the quote from their p. 50, given above). Third, that subjects in fact construed their task thus. The evidence for this, detailed above, is extremely weak.

Two things seem to have misled Wolford et al. (1990). First, when they attempted to construct a "more concrete and familiar scenario" (p. 48) for studying the conjunction fallacy, they inadvertently constructed one that, unlike the original Linda problem, really can be regarded as matching a fixed set of candidates to occupations.² Second, by thinking of the crucial factor as consisting of whether or not the outcomes were "known," they made the erroneous inferential step that since in the Linda problem the outcomes can be viewed as either "known" or "unknown," the task can also be construed according to the "known model" or the "unknown model," respectively.

Many writers have attempted to defend seemingly erroneous responses by offering interpretations of subjects' reasoning that rationalizes their responses. Sometimes, however, this charitable approach has been misguided, either because the subjects are quick to acknowledge their error themselves once it is pointed out to them, or because the interpretation required to justify the response is even more embarrassing than the error it seeks to excuse.³ In the present case, there is evidence that statistically sophisticated subjects would not defend the conjunction fallacy. Kahneman and Tversky (1982) allowed subjects who were given the Linda problem to choose between the following two arguments:

- (1) [Bank teller is more probable than feminist bank teller] because the probability that Linda is both a bank teller and an active feminist must be smaller than the probability that she is a bank teller.

(2) [Feminist bank teller is more probable than bank teller] because Linda resembles a bank teller who is active in the feminist movement more than she resembles a bank teller. (p. 496)

The first argument was endorsed by 83% of these subjects, who "immediately recognized the validity of the conjunction rule" (p. 497).

Let us now leave aside the normative issue of which model is appropriate for the Linda problem, as well as the empirical issue of which model is in subjects' minds. As students of cognition, let us now ask whether Wolford et al.'s (1990) results have taught us anything new about prediction under uncertainty.

The focus on biases and fallacies is popular in the study of judgment under uncertainty. It derives its methodological power from its diagnosticity regarding underlying cognitive processes. When subjects judge the probability of a conjunction of a representative event with an unrepresentative one to lie between the probabilities of the component events, we have nice evidence that they are not using the logic of probability, according to which these probabilities can never be thus ordered, but rather the logic of similarity, which predicts just such an order. But when subjects make a judgment that is compatible with both the laws of probability theory and the representativeness heuristic, their responses are useless for distinguishing between the two.

Assume for the sake of argument that subjects given the Linda problem really were judging $P(L|F\&B)$ versus $P(L|B)$, as claimed by Wolford et al. (1990), and not $P(F\&B|L)$ versus $P(B|L)$, as Tversky and Kahneman's (1983) analysis assumes. For these subjects, an account by representativeness is unnecessary. It is sufficient to argue, as Wolford et al. (1990) do, that "people intuitively prefer a Bayesian approach to these problems" (p. 53, n. 3). On the other hand, these subjects' ranking of the outcomes—even if it is not fallacious—is nonetheless still perfectly compatible with the representativeness heuristic. Hence, arguing that people are normative Bayesian reasoners is unnecessary. It is sufficient to argue from representativeness. If subjects' ordering of outcomes for Linda can be explained both as normative Bayesian reasoning and as heuristic reasoning, which account is preferable? Wolford et al. concede that (1) "many [indeed, most] subjects in the most favorable conditions continue to make the conjunction fallacy" (p. 53; see also Table 2, p. 52); (2) their claim for Bayesianism is "at odds with considerable research showing that people are 'bad Bayesians'" (p. 53, n. 3); and (3) people invoke the Bayesian model "even when that model is inappropriate" (p. 53). In the face of these difficulties with the Bayesian account, it seems more parsimonious to remain with the standard representativeness account for the Linda problem results.

A question may remain as to how the representativeness heuristic could account for the full range of findings reported

by Wolford et al. (1990), and in particular, for the varying rates of the conjunction fallacy across various kinds of problems and subjects. Note that Tversky and Kahneman (1983) never said that all of the people use representativeness all of the time. Indeed, they were surprised that even statistically sophisticated subjects order outcomes by representativeness, and that this happens even in "transparent tests" (those in which the only options to be ranked are B and $F\&B$). All previous research reports a minority, of varying size, that does not commit the conjunction fallacy. Why? Who knows. Perhaps their judgments of representativeness differ from the majority; perhaps they realize the logical necessity of the conjunction rule; perhaps they are guessing carelessly, etc. It is quite possible that various cues and features of problems encourage the use of the representativeness heuristic more or less. It is even possible that one such feature is whether the outcomes are "known" or "unknown." However, saying that the rates of usage of the heuristic fluctuate somewhat with different features of problems and subjects is not the same as saying what Wolford et al. (1990) seem to be saying: that in certain problems, subjects by and large use normative considerations rather than heuristic thinking.

REFERENCES

- BAR-HILLEL, M. (1984). Representativeness and probability judgments. *Acta Psychologica*, 55, 91-107.
- COHEN, L. J. (1979). On the psychology of prediction: Whose is the fallacy? *Cognition*, 7, 385-407.
- EDDY, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249-267). Cambridge, U.K.: Cambridge University Press.
- KAHNEMAN, D., & TVERSKY, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430-454.
- KAHNEMAN, D., & TVERSKY, A. (1982). On the study of statistical intuitions. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 493-508). Cambridge, U.K.: Cambridge University Press.
- TVERSKY, A., & KAHNEMAN, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293-315.
- WOLFORD, G., TAYLOR, H. A., & BECK, J. R. (1990). The conjunction fallacy? *Memory & Cognition*, 18, 47-53.

NOTES

1. Here and throughout, we assume that these conditionals are well defined.
2. In that problem the candidates were students, and the "occupations" were class schedules.
3. For example, Cohen (1979) defended the finding that people judge the probability of days with over 60% male births to be equally likely in large and in small hospitals (Kahneman & Tversky, 1972) by attributing to respondents the belief that obstetricians in those hospitals practice infanticide.