Adams's Puzzle about Counterfactuals
Dorothy Edgington Konstanz 04.09.2010 Handout

1. Prologue. Some highlights of Adams's philosophy

(a) His probabilistic criterion of validity: an argument is valid iff the uncertainty (1 – the probability) of the conclusion cannot exceed the sum of the uncertainties of the premises; or equivalently: for any $\delta > 0$ there is an $\varepsilon > 0$ such that if all the premises have at least $1-\varepsilon$, the conclusion has at least $1-\delta$.

(b) He was first to develop a logic with the features later incorporated into the possible-world semantics of Stalnaker and Lewis. He was first to give counterexamples to strengthening, contraposition and transitivity. E.g. 'If Brown wins the election, smith will resign immediately afterwards; if Smith dies before the election, Brown will win the election; so, if Smith dies before the election, Smith will resign immediately after the election'. Adams 1965, p. 166. These sentences can consistently be assigned $1-\varepsilon$, 1 and 0 respectively.

2. Some background and *some* motivation for thinking of conditional judgements in terms of conditional probabilities. We need to take as of central importance the fact that the non-trivial contingent judgements we make are typically uncertain judgements.

3. Adams on counterfactuals and their connection with Bayesian reasoning.

Informal examples:

You are driving, of an evening, in the dark, close to the house of some friends, and have considered paying a visit. You turn the corner. 'They're not at home', you say, 'for the lights are off. And if they had been at home the lights would have been on'.

A patient is brought to the hospital in a coma. 'I think he must have taken arsenic', says the doctor, after examination, 'for he has [such-and-such] symptoms. And these are just the symptoms he would have if he had taken arsenic'.

Bayes's Theorem

$$\frac{p_O(H)}{p_O(\neg H)} \times \frac{p_O(E \mid H)}{p_O(E \mid \neg H)} = \frac{p_O(H \mid E)}{p_O(\neg H \mid E)} = \frac{p_N(H)}{p_N(\neg H)}$$

4. Adams's Puzzle (*The Logic of Conditionals* p. 129)

"Imagine the following situation. We have just entered a room and are standing in front of a metal box with two buttons marked 'A' and 'B' and a light, which is off at the moment, on its front panel. Concerning the light we know the following. It may go on in one minute, and whether it does or not depends on what combinations of buttons A and B, if either, have been pushed a short while before, prior to our entering the room. If exactly one of the buttons has been pushed then the light will go on, but if either both buttons or neither has been pushed then it will stay off. We think

it highly unlikely that either button has been pushed, but if either or both were pushed they were pushed independently, the chances of A's having been pushed being 1 in a thousand, while the chance of B's having been pushed is a very remote 1 in a million. In the circumstances we think there is only a very small chance of 1,000,999 in one billion (only very slightly above 1 in a thousand) that the light will go on, but a high probability of 999 in a thousand that *if B was pushed the light will go on*.

Now suppose to our surprise that the light does go on, and consider what we should infer in consequence. Leaving out numerical probabilities for the moment, we would no doubt conclude that the light probably lit because A was pushed and B wasn't, and not because B was pushed and A wasn't (the former being about 1000 times more likely than the latter). Therefore, since A was probably the button pushed, *if B had been pushed the light would not have gone on*, for then both buttons would have been pushed. The point is that the counterfactual would be affirmed [now] despite the fact that the corresponding indicative was very improbable [earlier], because its contrary 'if B was pushed then the light will go on' had a prior probability of 0.999."

Adams's formalisation of the above reasoning (up to a certain point)

$$\frac{p_N(B)}{p_N(A)} = \frac{p_O(B)}{p_O(A)} \times \frac{p_O(L \mid B)}{p_O(L \mid A)}$$

$po(L \text{ given } B) = po(\neg A) = 0.999$

$po(L \text{ given } A) = p_O(\neg B) = 0.999999$

So

$$\frac{p_N(B)}{p_N(A)} = \frac{0.00001}{0.001} \times \frac{0.999}{0.999999} = \frac{999}{999999}$$

Adams proves that in any probability distribution whatsoever

$p(\neg L \mid B) = p(A \mid B)$

5. The 'Law of Total Conditional Probability'; or, the battle between two formulas.

$p(A) = p(A\&X) + p(A\&\neg X) = p(A \mid X).p(X) + p(A \mid \neg X).p(\neg X).$

This is an instance of what is sometimes called the law of total probability.

For instance you think it's 50-50 whether bag X or bag Y is in front of you. (So Y is in effect ¬X.) Each bag contains 100 balls. A ball is to be selected at random. In bag

X (you know) 90% of the balls are red. In bag Y, 10% of the balls are red. How likely is it that you will pick a red ball? (90% × 50%) + (10% × 50%) = 50%. (Of course.)

Now consider the conditional case. You're wondering how likely it is that C if A. It depends on whether X or ¬X. You know the chance of C if A&X; and you know the chance of C if A&¬X. How should you proceed?

First guess: (*) $p(C \mid A) = p(C \mid A\&X).p(X) + p(C \mid A\&\neg X).p(\neg X)$.

For instance, again you think it's 50-50 whether bag X or bag Y is in front of you. (So Y is in effect ¬X.) In bag X, 90% of the red balls have black spots. In bag Y, 10% of the red balls have black spots. How likely is it that if you pick a red ball it will have a black spot? According to the above formula: (90% × 50%) + (10% × 50%) = 50%.

But wait! These bags, X and Y, are the same bags as before. There's a far higher proportion of red balls in bag X than in bag Y. So if I pick a red ball, that makes it likely that it's bag X in front of me, in which case it's likely to have a black spot. Thinking this way, it's well over 50% likely that if I pick a red ball, it will have a black spot.

In fact (*) is not a theorem. Instead we can derive:

(**) $p(C \mid A) = p(C \mid A\&X).p(X \mid A) + p(C \mid A\&\neg X).p(\neg X \mid A)$.

Proof: $$p(C \mid A) = \frac{p(C\&A)}{p(A)} = \frac{p(C\&A\&X) + p(C\&A\&\neg X)}{p(A)}$$

$$= \frac{p(C \mid A\&X).p(A\&X) + p(C \mid A\&\neg X).p(A\&\neg X)}{p(A)}$$

$$= p(C \mid A\&X).p(X \mid A) + p(C \mid A\&\neg X).p(\neg X \mid A).$$

Applying this to the example, we get (90% × 90%) + (10% × 10%) = 82%..

Now, the probability functions may represent credence, or they may represent objective chances. Or they may be a mixture generated by something like what Lewis called the principal principle. In the unconditional case: Suppose I have (keeping things simple) two exclusive and exhaustive hypotheses H_1 and H_2, about the chance of A. Then the probability I should assign to A is $ch(A \mid H_1).p(H_1) + ch(A \mid H_2).p(H_2)$.

In the conditional case, suppose I have two exclusive and exhaustive hypotheses, H_1 and H_2, about the chance of C given A. It might be tempting to think that

$p(C \mid A) = ch(C \mid A\&H_1).p(H_1) + ch(C \mid A\&H_2).p(H_2)$. But that goes with (*) which is not derivable as a theorem. Instead we should have, as the analogue for conditional probabilities:

$p(C \mid A) = ch(C \mid (A\&H_1).p(H_1 \mid A) + ch(C \mid A\&H_2).p(H_2 \mid A)$.

Douven's tweaking of the balls example. Bag Y contains 100 balls, 10 of them red, 1 of the red ones with a black spot. Bag Y is wearing thin, so someone puts inside it two bags, Y_1 and Y_2, and distributes the contents between them. Each ball in Y still, as before, has an equal chance of being picked, so this little cosmetic alteration should make no difference. Y_1 and Y_2 each contain 50 balls. Y_1 contains 1 red ball, which has a red spot. Y_2 contains 9 red balls, none with a black spot.
Applying (*) before this alteration gave the answer 0.5. Applying (*) after the alteration gives:

$p(B \mid R) = p(B \mid R\&X).p(X) + p(B \mid R\&Y_1).p(Y_1) + p(B \mid R\&Y_2).p(Y_2)$

$= (0.9 \times 0.5) + (1 \times 0.25) + (0 \times 0.25) = 0.7$.

Using instead the kosher principle (**) we get

$p(B \mid R) = p(B \mid R\&X).p(X \mid R) + p(B \mid R\&Y_1).p(Y_1 \mid R) + p(B \mid R\&Y_2).p(Y_2 \mid R)$

$= (0.9 \times 0.9) + (1 \times 0.1) + (0 \times 009) = 0.82$, as before.

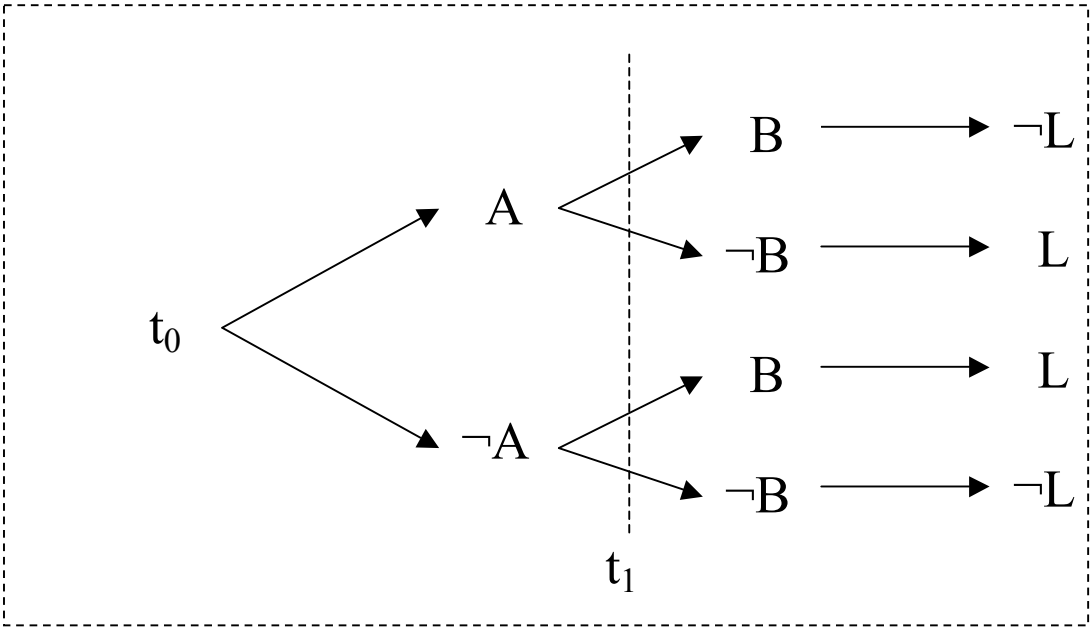
6. Back to Adams, and Skyrms

Adams's suggestion: $p(B \rightarrow \neg L) = \sum (i = 1, \ldots n) p_i(S_i) p_0(B\&S_i \rightarrow \neg L)$.

Skyrms: The 'Basic Assertability Value' of the counterfactual $p \rightarrow q$ is the weighted sum of the various possible hypothetical objective chances of q given p, the weights being your latest epistemic probability for those hypotheses.

7. More switching cases. E.g. 'I I had bet on heads I would have won'. I argue that the ultimate value to be assigned to 'If A had been the case, C would have been the case is not necessarily the chance, back then, of C given A, but the chance, back then, of C given A&S where S includes all subsequent facts, casually independent of A, which have some bearing on whether C. It is still a conditional chance, but not just of C given A!

8. An attempt to solve Adams's puzzle.



Adams missed this possibility: the appropriate past hypothetical standpoint for evaluating 'If B had been pushed the light would not have gone on', is of someone at t_1. A is already settled and they think it's around 99.9% likely that A. B is still a chance event in the future, independent of A. They know that $ch(\neg L \mid B) = 1$ if A, and 0 if ¬A. Because of the independence they can us the 'bad' formula: $p(\neg L \mid B) = p(\neg L \mid B\&A).p(A) + p(\neg L \mid B\&\neg A).p(\neg A) \approx (1 \times 0.999) + (0 \times 0.001) \approx 0.999$.

Some references:
Adams 1965, 'A Logic of Conditionals'. *Inquiry*.
Adams 1966, 'Probability and the Logic of Conditionals', in Hintikka and Suppes (eds), *Aspects of Inductive Logic*. Reidel.
Adams 1975: *The Logic of Conditionals*. Reidel.
Douven 2008: 'Kaufmann on the Probabilities of Conditionals'. *Journal of Philosophical Logic*.
Edgington 1978, review of Adams, *The Logic of Conditionals*. *Mind*.
Edgington 2004, 'Counterfactuals and the Benefit of Hindsight', in Dowe and Noordhof (eds) *Cause and Chance*
Kaufmann 2004, 'Conditioning against the Grain', *Journal of Philosophical Logic.*
Skyrms 1981, 'The Prior Propensity Account of Subjunctive Conditionals' in Harper, Stalnaker and Pearce (eds), *Ifs*. Reidel.