Episteme, 17, 1 (2020) 64–72 © Cambridge University Press
doi:10.1017/epi.2018.17

A PUZZLE ABOUT EXPERTS, EVIDENTIAL SCREENING-OFF AND CONDITIONALIZATION

ITTAY NISSAN-ROZEN
ittay.nissan@mail.huji.ac.il

ABSTRACT

I present a puzzle about the epistemic role beliefs about experts' beliefs play in a rational agent's system of beliefs. It is shown that accepting the claim that an expert's degree of belief in a proposition, A, screens off the evidential support another proposition, B, gives to A in case the expert knows and is certain about whether B is true, leads in some cases to highly unintuitive conclusions. I suggest a solution to the puzzle according to which evidential screening off is rejected, but show that the price of this solution is either giving up on the mere idea of deferring to expert's opinion or giving up on Bayesian conditionalization.

1. THE PUZZLE

Consider the following three propositions.

H – John is a physics professor at Harvard
T – John has published at least 3 papers in physics journals
X – Jane believes A to degree x
(when x is a number in the unit interval).

Assume that although you are uncertain about whether T is true, you are certain that Jane knows and is certain about *whether* T is true. In addition you are certain that Jane has much more information about John (who you have never met and know almost nothing about) than you and you are certain that she is more epistemically competent than you in evaluating this evidence (you can assume any level of epistemic competence that you would like).

In other words, when it comes to proposition H, Jane is both what Hall (2004) calls a "database-expert" and what he calls an "analyst-expert". She has all the information you have and more information and she is better than you in evaluating it. Furthermore you are certain that this is the case. It seems reasonable, in such a case, to adopt the following conditions on your credence function:

Epistemic Deference (ED): $c(H|X) = x$
Strong Epistemic Deference (SED): $c(H|XT) = x$

ED seems reasonable because Jane has more relevant information than you and she is more epistemically competent than you, so you should defer to her judgement about

https://doi.org/10.1017/epi.2018.17 Published online by Cambridge University Press

the subject matter. SED seems reasonable because, on the assumption that Jane knows and is certain about whether T is true, her judgement about the truth of H screens off the evidential support T gives to H.

Another way to put the point is this: T is either true or false. In case T is true, Jane already took this fact under account when forming her belief about H and so taking T to be indicative with respect to H over and above X would constitute double counting. In case T is false, you surely should just ignore the evidential support T would give to H if it were true and adopt Jane's degree of belief in H (which was formed under the assumption that T is false).[1]

From ED and SED we get:

Screening (S): c(H|X) = c(H|XT)

However, assuming c(HX) ≠ 0, it is straightforward to show that S holds iff Reverse Screening (RS) holds:

Reverse Screening: (RS): c(T|HX) = c(T|X)

$$\text{(Proof : c(T|HX)} = \frac{c(HTX)}{c(HX)} = \frac{c(H|XT)c(T|X)}{c(H|X)}\text{, so c(T|HX)} = \text{c(T|X) iff c(H|XT)} = \text{c(H|X))}$$

Now suppose you become certain that Jane believes H to a low though positive degree (maybe she tells you "I am almost certain but not 100% certain that John is not a professor"). Then, assuming ED, for any X to which you assign a positive probability it must be true that c(HX) ≠ 0. Thus, RS holds with respect to every X to which you assign a positive probability.

However, c(T|HX) is very high (if John is a physics professor at Harvard he probably has published at least 3 papers in physics journals) so it follows from RS that c(T|X) is very high for every X to which you assign a positive probability. Thus c(T) must be very high,

1 There seems to be a wide agreement among philosophers that ED must be respected by a rational agent in the type of cases I have described. See for example Hall (2004), Elga (2007), Joyce (2007). Here I do not assume that ED is a condition of rationality. I only assume that a rational agent is *allowed* to respect it. My discussion is also limited to cases of deference to human experts. As an anonymous referee pointed to me, when considering ED in cases in which the expert is ideally rational and maximally informed (maybe chance – as Hall (2004) argues – is such an expert) our intuitions regarding formally equivalent examples dramatically change (I thank a referee for this and other extremely useful comments). SED is usually discussed only in the context of cases in which the expert/peer (the role that Jane plays in our example) does not know whether T is true. In such cases SED should clearly be rejected. However, when you are certain that Jane knows (and is certain about) whether T is true SED seems, at first sight, highly intuitive. Furthermore and as will be discussed below, an agent that respects ED but violates SED in the type of cases discussed, must violate ED after Bayesian conditionalization on T. Thus, without giving up on Bayesian conditionalization, it is impossible to violate SED without violating ED at some point (in which all the conditions of the example hold, i.e. when Jane has more information than you and is more epistemically competent than you).

https://doi.org/10.1017/epi.2018.17 Published online by Cambridge University Press

but this is absurd since the only piece of information you have which seems relevant to T is that Jane is almost certain that H is false!

2. AN UNSUCCESSFUL SOLUTION

One might be tempted to offer the following solution to the puzzle. Although c(T|H) is very high, at least with regard to some values of x, c(T|HX) can be low. If this is true then, granted that you assign high enough probabilities to the relevant hypothesis about Jane's degree of belief in H, c(T) can still be low.

The thought might be that since Jane knows whether T is true (i.e. she knows whether John has published at least 3 papers in physics journals) the fact that she believes H to a given low degree is (for some low degrees) indicative for the falsity of T. This is so because she might believe H to a low degree *because* (or partly because) she knows that T is false (i.e. she might believe to a high degree that John is not a physics professor at Harvard because she knows he has not published at least 3 papers in physics journals).

While I agree that this line of reasoning can support the claim that c(T|HX) might be lower than c(T|H), I do not see how it can plausibly support the claim that c(T|HX) might be low in absolute terms (say lower than 0.5). Jane might believe H to a low degree for many reasons. Jane might know that John lives in London or she might know that he once took an oath never to teach in Harvard, or she might know that he spends most of his waking hours playing soccer. Each one of these possibilities is highly indicative to the falsity of H but consistent with H and T being true. Thus, each one of these possibilities can plausibly explain Jane's low degree of belief in H. I see no reason to assume that a rational agent must put so much weight on the falsity of T as a possible explanation for Jane's low degree of belief in H so that c(T|HX) will turn out to be low.[2]

To get a sense of how implausible the claim that c(T|HX) must be low with respect to enough values of x is, consider the following Lemma:

Lemma 1: c(T|HX) < 0.5 iff c(−TX|H) > c(TX|H)
Proof:

$$c(T|HX) = \frac{c(THX)}{c(HX)} = \frac{c(TX|H)}{c(X|H)} = \frac{c(TX|H)}{\frac{c(THX)+c(-THX)}{c(H)}} = \frac{c(TX|H)}{c(TX|H)+c(-TX|H)}$$
$$= \frac{1}{1+\frac{c(-TX|H)}{c(TX|H)}}.$$

Now, since $\frac{c(-TX|H)}{c(TX|H)} > 1$ *iff* $c(-TX|H) > c(TX|H)$, $c(T|HX) > 0.5$ *iff* $c(-TX|H) > c(TX|H)$.

2 I do not doubt that it is possible to tell some background stories that will make it plausible to claim that c(T|HX) is low. However, in order to use the claim that c(T|HX) is low to resolve the puzzle, one must argue that it is always the case that there are enough values of x with respect to which c(T|HX) is low. This latter claim is the one that I find implausible.

https://doi.org/10.1017/epi.2018.17 Published online by Cambridge University Press

The condition c(-TX|H) > c(TX|H) states that under the assumption that John is a physics professor at Harvard, it is more likely that John has not published at least 3 papers in physics journals and that Jane believes H to degree x than that John has published at least 3 papers in physics journals and that Jane believes H to degree x. However, since John being a physics professor at Harvard is highly indicative for John publishing more than 3 papers in physics journals, this conditions seems to be – in most cases – false.

An anonymous referee presented the following objection.

$$c(-TX|H) > c(TX|H) \text{is true iff } c(-TH|X) > c(TH|X)$$

$$\text{(since } c(-TX|H) = \frac{C(-THX)}{C(H)} = \frac{C(-TH|X)c(X)}{C(H)} \text{ and } c(TX|H) = \frac{C(THX)}{C(H)} = \frac{C(TH|X)c(X)}{C(H)}\text{)}.$$

However, c(−TH|X) > c(TH|X) does intuitively seem true in case x is low. Thus, it might be the case that it is indeed true that there are always enough values of x (i.e. enough propositions of the form X to which you assign a positive probability) with respect to which c (T|HX) < 0.5 so that c(T) can be low.

I think this is an interesting objection that sheds new light on the significance of lemma 1, but I believe it ultimately fails. Here is why.

In order for this objection to solve the puzzle, it must be the case that

$$c(-TH|X) > c(TH|X) \qquad 1$$

is always a plausible condition; i.e. that there are no cases in which 1 is an implausible demand regarding all the values of x (Jane's degree of belief in H) to which you assign a positive credence. If there are some cases in which 1 is implausible with respect to all (or enough, but I can do with "all") values of x, the puzzle still holds in those cases.

Now while I agree that one can build cases in which 1 seems plausible with respect to some values of x, I am convinced that there are many cases in which 1 is implausible with respect to all values of x.

To see that, notice that on the face of it, the proposition HT is much more probable than the proposition H-T. This is uncontroversial. The objection is that learning a X with a low x changes things because the fact that T is false (the fact that John has not published at least 3 papers) *explains* Jane's high confidence that John is not a Harvard professor (even though he is) and so under the assumption that Jane believes H to a low degree, H-T is more probable than HT.

However, and as mentioned above, this consideration can support condition 1 only *in the absence of alternative and more plausible explanations for Jane's low degree of belief in H*. However, often there are many such alternative and more plausible explanations for Jane's low degree of belief in H.

To demonstrate this, let us just make one such an explanation explicit:

M – "John is a history professor at MIT"

https://doi.org/10.1017/epi.2018.17 Published online by Cambridge University Press

Suppose you are certain that Jane knows and is certain about whether M is true but you are uncertain whether it is true. Still you are almost certain that it is true. So, just to make things clear, you are almost (but not entirely) certain that John is a MIT history professor. You do not rule out the possibility that John is a physics professor at Harvard, but you think it is very unlikely (because you are almost certain that he is a history professor at MIT) and you do not rule out the possibility that John has both a physics position at Harvard and a history position at MIT, but naturally you also take this to be very unlikely.

The puzzle, recall, is that, assuming ED and SED, you must be almost certain that John has published at least 3 physics papers and this seems absurd because all the information you have about John which seems relevant is that Jane is almost certain that he is not a physics professor at Harvard.

Now, the proposed solution is to avoid the puzzle by arguing that for enough values of x, c(−TH|X) > c(TH|X).

However:

$$c(\mathrm{TH}|\mathrm{X}) = \frac{C(THX)}{C(X)} = \frac{C(THXM)}{C(X)} + \frac{C(THX-M)}{C(X)} = c(\mathrm{THM}|\mathrm{X}) + c(\mathrm{TH}-\mathrm{M}|\mathrm{X}) \quad 2$$

and similarly

$$c(-\mathrm{TH}|\mathrm{X}) = c(-\mathrm{THM}|\mathrm{X}) + c(-\mathrm{TH}-\mathrm{M}|\mathrm{X}) \quad 3$$

and so, since the truth of M (together with your certainty that Jane knows and is certain about whether M is true) also explains Jane's low degree of belief in H, it seems that:

$$c(\mathrm{THM}|\mathrm{X}) \text{ should be higher than } c(-\mathrm{THM}|\mathrm{X}). \quad 4$$

To see that, suppose you have learnt that Jane actually believes H to a given low degree (this is unsurprising for you, of course, since you are certain Jane is an expert and you are almost certain that John is a MIT history professor). Now, which one of the two options seems more probable to you, that John holds both positions without publishing 3 papers or that John holds both positions after publishing at least 3 papers (both of them seems very unlikely but the question is which one seems more unlikely)? It is obvious, I take it, that you should take the latter to be more probable.

Now, it might indeed be true (in line with the objection) that:

$$c(-\mathrm{TH}-\mathrm{M}|\mathrm{X}) > c(\mathrm{TH}-\mathrm{M}|\mathrm{X}) \text{ (for the reasons outlined above)}. \quad 5$$

However, by assumption c(−TH−M|X) is much lower than c(−THM|X) and c(TH−M|X) is much lower than c(THM|X) (because you are almost certain that M is true) so the difference between (THM|X) to (−THM|X) can easily outweigh the difference between c (−TH−M|X) to c(TH−M|X) and thus in many cases it must be that c(TH|X) > c(−TH|X).

More generally, although it is indeed the case that *ceteris paribus* learning that Jane believes H to a low degree gives evidence in support of −TH when compared to TH, because TH is initially much more probable than −TH and because there can be more evidence that undercuts the evidential support that learning that Jane believes H to a

https://doi.org/10.1017/epi.2018.17 Published online by Cambridge University Press

low degree gives to −TH relative to TH, there seems to be no reason to assume that it must always be the case that c(−TH|X) > c(TH|X) for enough values of x.

Here is another way to make the point:

Surely, for a low x:

$$c(X|H-T) > c(X|HT). \quad 6$$

6, I take it, accurately expresses the intuition that the falsity of T can explain Jane's low degree of belief in H (even though H is true) and is uncontroversial. However 6 does not imply 1. We can rewrite 6 in the following way:

$$\frac{c(-TH|X)c(X)}{c(H-T)} > \frac{c(TH|X)c(X)}{c(HT)} \quad 7$$

and thus 6 is true iff:

$$\frac{c(-TH|X)}{c(HT|X)} > \frac{c(-TH)}{c(TH)} \quad 8$$

Notice that 8 can be true while c(TH|X) is higher than c(−TH|X) (i.e. while 1 is false), as long as c(TH) is enough higher than c(−TH). Since in many cases c(TH) is indeed much higher than c(−TH) and since the existence of alternative explanations for Jane's low degree of belief in H reduces the difference between c(TH|X) and c(−TH|X), there must be many cases in which 1 should be rejected with respect to all values of x to which you assign a positive probability.

Thus, in many scenarios, the claim that c(T|HX) is high should not be rejected and so the puzzle stands.

3. A SUCCESSFUL SOLUTION: REJECTING SED

Since the conclusion that c(T) is very high cannot be the right conclusion to draw, the problem must lay, I believe, in Screening. In order to reject S we must reject either ED or SED and indeed a closer examination of SED may lead one to reject it.

In a nutshell the problem is this: regarding any X, from your point of view it might be true when T is true and it might be true when T is false. Consider now an X with a low x. If X is true when T is false then X is true *partly because* T is false (as the falsity of T is an indication for the falsity of H and so Jane's low degree of belief in H must be sensitive to that). If X is true when T is true then X is true *in spite of* T being true (as the truth of T is a weak indication for the truth of H and so Jane's degree of belief in H must be sensitive to that). Thus, under the assumption that X is true (that Jane believes H to a given low degree), the other evidence available to Jane (i.e. not including whether T is true) is more indicative of H in case T is true, than in case T is false.

Thus, c(H|TX) should intuitively be higher than c(H|−TX) since Jane believing H to degree x *even though* T is true is more indicative to the truth of H than Jane believing H to degree x *partly because* T is false. However if c(H|TX) > c(H|−TX) Screening must be false.

https://doi.org/10.1017/epi.2018.17 Published online by Cambridge University Press

To make the point clearer consider a case with the same formal structure, only one in which T follows from H. For example consider the following three propositions.

N – John never got married
B – John did not get married until 2014
Y – Jane believes N to degree y
(when y is the a number in the unit interval).

As before, suppose you are uncertain about whether B is true but you are certain that Jane knows and is certain about whether B is true. In addition you are certain that Jane has much more information relevant for the assessment of N than you and that she is more epistemically competent than you.

In this example, unlike in the previous one, it seems reasonable to demand that for any $y \neq 0$, $c(-BY) = 0$. Why? Because assuming that Jane is minimally epistemically competent (and we have stipulated that you are certain that this is the case), in case B is false, Jane believes that N is true to degree 0 (because if she is certain that John got married before 2014, she is certain that it is not true that he never got married). Thus, you must be certain that it cannot be the case that B is false and Jane believes N to a degree other than 0. Now, if $c(-BY) = 0$, $c(N|-BY)$ is undefined and for any Y such that $y \neq 0$, S trivially holds (since BY follows from Y).

What this case shows is that the truth of Y bears on N in a completely different way when B is true and when B is false. When B is true all goes well: Jane forms her degree of belief partly on the basis of B and since you want to avoid double counting you just adopt her degree of belief in N. However, when B is false your assumption regarding Jane's epistemic competence is inconsistent with Jane holding any non-zero degree of belief in N.

Coming back to our original example in which T does not follow from H but only becomes highly probable on the assumption that H, we should draw an analogous conclusion: the truth of X bears on H differently when T is true and when T is false. Jane needs less independent (that is independent of T) evidential support for H in order for her to adopt a degree of belief x in H, in case T is true than in case T is false. Thus $c(H|TX)$ should not be equal to $c(H|-TX)$ and thus S cannot be true.

Another way to put the same point is the following one. Learning TX teaches you something about H not only directly – i.e. by way of the evidential support T and X gives to H – but also indirectly – i.e. by teaching you something about Jane's evidential basis for adopting the degree of belief she has adopted in H. Specifically, it teaches you that the evidence available to Jane regarding the truth of H, not including T, is relatively weak (that is relative to the other possibility you consider: that X is true while –Y is true).

If this is right then we have a good reason to reject SED and thus also S and the implausible conclusion that c(T) must be high can be avoided.

4. THE PRICE: EITHER REJECTING ED OR REJECTING BAYESIAN CONDITIONALIZATION

Suppose the conclusion of the previous section is the correct one to draw and we reject SED. In such a case

$$c(H|TX) \neq c(H|X).$$

https://doi.org/10.1017/epi.2018.17 Published online by Cambridge University Press

Now you learn that T is true. What is your conditional degree of belief in H given X? Let c'(.) be your credence function after learning T. If you update your beliefs in a way consistent with Bayesian conditionalization then:

$$c'(\mathrm{H}|\mathrm{X}) = \frac{c'(HX)}{c'(X)} = \frac{c(HX|T)}{c(X|T)} = c(\mathrm{H}|\mathrm{TX}) \neq c(\mathrm{H}|\mathrm{X}) = x.$$

Thus, if before learning T you have obeyed ED, after learning T and updating using Bayesian conditionalization you no longer obey ED.

This conclusion is, however, highly unintuitive. By stipulation, even after learning T, you are still certain that Jane is more epistemically competent than you and has much more evidence relevant for the assessment of H than you, so any argument in favour of accepting ED before learning T holds also for accepting ED after learning T.

Notice also that the conclusion at which we have arrived is not only that ED cannot be a condition of rationality both before and after conditionalization. The conclusion is much stronger than that: a rational agent *must* violate ED either before or after conditionalization. Accepting this conclusion seems to me like a very high price to pay. Even if it is not irrational not to adopt the degree of belief of an expert that also have much more evidence than you, it is surely not irrational to adopt it.

The other alternative is to reject both SED and ED while also rejecting the claim that after learning X (i.e. after learning the expert's opinion) a rational agent must update her beliefs in a way which is inconsistent with Bayesian conditionalization. Instead she should just adopt the expert's degree of belief. This solution manages to avoid the conclusion that c(T) must be high while capturing – in a different way than a commitment to ED does – the intuition that in some cases rational agents should defer to experts' opinions. The thought here might be that while it is true that after learning the degree of belief of an expert in H, a rational agent must (or at least can) adopt it, when considering hypotheses about what this degree of belief is, the evidential support each such hypothesis gives to H is different than the effect learning that this hypothesis is true will have on the agent's degree of belief in H.

I do not find this solution particularly attractive (due to the fact that – other than for the purpose of avoiding the puzzle – its relaxation of Bayesian conditionalization seems unmotivated). However, I do not see any other way to go. Further research is needed in order to determine whether this is the right conclusion to adopt.[3]

REFERENCES

Elga, A. 2007. 'Reflection and Disagreement.' *Noûs*, 41: 478–502.

3 This paper has emerged from an exercise on Bayesian epistemology I wrote for my 2016–17 Introduction to Epistemology course (15302) at the Hebrew University. I thank the students in the course for challenging me to find interesting ways to teach epistemology and by doing so to gain a better understanding of several key issues in epistemology. I especially thank Merav Kaplan, my TA in that course, for her deep and thoughtful comments on earlier versions of this paper. I also thank Dani November, David Enoch, Levi Spectre, Noam Nisan, Ruth Weintraub and two anonymous referees for their very helpful comments.

https://doi.org/10.1017/epi.2018.17 Published online by Cambridge University Press

Hall, N. 2004. 'Two Mistakes about Credence and Chance.' *Australasian Journal of Philosophy*, 82: 93–111.

Joyce, M. J. 2007. 'Epistemic Deference: The Case of Chance.' *Proceedings of the Aristotelian Society*, 107: 187–206.

ITTAY NISSAN-ROZEN is a Lecturer in the Department of Philosophy and the PEP program at the Hebrew University of Jerusalem. His research focuses on issues in epistemology, decision-theory, philosophy of science, ethics and political philosophy. He has published in *Philosophy and Phenomenological Research*, *Analysis*, *Synthese*, *British Journal for the Philosophy of Science, Economics and Philosophy* and more.

https://doi.org/10.1017/epi.2018.17 Published online by Cambridge University Press