ON A PUZZLE ABOUT EXPERTS, SCREENING-OFF AND THE RARITY OF DEFEAT

ABSTRACT. We introduce a "rarity of defeat" principle, valid in cases of deference to an expert, to address intuitions involved in a puzzle of Nissan-Rozen concerning epistemic deference and evidential screening-off.

Ittay Nissan-Rozen (2018) invites us to consider the following propositions.

H – John is a physics professor at Harvard
T – John has published at least 3 papers in physics journals
$X = X(x)$ – Jane believes H to degree x[1]

Nissan-Rozen stipulates that we know that Jane knows whether or not T is true, and that, moreover, "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...."

Nissan-Rozen claims (he cites Hall 2004, Elga 2007 and Joyce 2007 in support of this claim) that there is "widespread agreement" that in cases like this we should defer to the more knowledgeable party, i.e. adopt her credences as our own, should we learn them. That is, our credence function c ought to have the property that $c(H|X) = x$; we should treat Jane as an "expert", exhibiting what Nissan-Rozen calls *Epistemic Deference*. But if that's the case, our credence function ought to satisfy both $c(H|TX) = x$ and $c(H|\neg TX) = x$ as well; Nissan-Rozen calls this *Strong Epistemic Deference*. For even if we were to learn that T is the case (say), Jane would still be more knowledgeable and competent than we are, and X says that Jane's credence in H is x.[2] So we should still defer, and adopt Jane's credence in H as our own. That is, conditional on X we ought to view T and H as *independent*.

[1]We follow Nissan-Rozen in writing simply "X" rather than "$X(x)$", though the reader should bear the dependence of X on x continuously in mind.

[2]A referee wrote "(Nissan-Rozen) seems to think that if Jane is an expert about H for us then she should remain one even after we learn the truth of T. This is wrong. Experts need not remain experts conditional on new information." But they do remain experts conditional on information that's not new to *them*. (Provided we know that it's not, as is the case here.) So this is an apparent misunderstanding.

That independence, however, does not sit well with Nissan-Rozen. For note that, since $c(H|TX) = c(H|\neg TX) = c(H|X)$ is equivalent to $c(T|HX) = c(T|\neg HX) = c(T|X)$ and the former holds, the latter also holds. But imagine now that we learn the truth of X for some low value of x. Since our only interesting evidence for or against T is that Jane has the low credence x in H, it seems reasonable that we should come to have low credence in T as well, i.e. $c(T|X)$ should be low. By independence, though, $c(T|HX) = c(T|X)$, so $c(T|HX)$ should be low. Nissan-Rozen rejects this conclusion out of hand, writing "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)".

Though the argument postures as a *reductio* of *Epistemic Deference*, there are reasons not to grant it this status. First there is some confusion caused by an inconsistency. It is stated at the beginning of the thought experiment that one can assume of the expert Jane "any level of epistemic competence that you would like". That can't be right, though, for that wouldn't preclude Jane being ideally rational, and that one ought to defer to ideally rationally agents who know more than we do is wholly unassailable. Indeed, there is cause to suspect that a referee for the paper called attention to this fact. Consider the following passage from footnote 1:

> My discussion is...limited to cases of deference to human experts. As an anonymous referee pointed to me [*sic*], when considering (*Epistemic Deference*) in cases in which the expert is ideally rational...our intuitions regarding formally equivalent examples dramatically change.

What the referee called attention to, we suspect, is that while the intuition that "$c(T|HX)$ is very high" might be appropriate to a case in which Jane is human (and therefore an imperfect reasoner), it isn't appropriate to a case in which Jane is ideally rational. If that's right then it was an oversight to fail to remove the earlier claim that one can assume of Jane "any level of epistemic competence that you would like". For in order for this argument to work, we have to believe that Jane might be imperfect.

That clears up one confusion, but it creates a new problem. For although there is surely "widespread agreement" that one ought to defer completely to the credences of a better-informed party if one is certain that she is ideally rational, there's no such agreement to the effect that one ought to defer completely to the credences of a better-informed party simply because one is "certain that she is more epistemically

competent". Indeed, without some reworking this criterion is wholly implausible, for it admits extreme cases in which we plainly should not defer, such as when we and the better-informed party exhibit different positive degrees of *anti-expertise*. (And, of course, the miniscule probability of $\neg TH$ relative to TH makes $c(T|HX)$ extremely sensitive to gratuitous reports of credence x on Jane's part in TH scenarios, so not even small deviations from "ideal" behavior can be safely ignored.)

In particular, we see little sympathy in Nissan-Rozen's sources for complete deference to non-ideal agents. Hall (2004) speaks of agents whose conditional probabilities are deferred to (he calls these "analyst-experts") being perceived as having "astonishing powers of evidential reasoning". Elga (2007) does write in one spot that "When it comes to the weather, I completely defer to the opinions of my local weather forecaster", but later he concedes that this isn't really true: "But upon finding out that my forecaster is confident that it will rain eggplants tomorrow, I will not follow suit. I will conclude that my forecaster is crazy." Thus sobered, he concludes that "only in highly idealized cases is it appropriate to treat someone as an expert..." Joyce (2007) meanwhile describes a range of partial deference scenarios: "C might see q's values as a better guide to A's truth-value than her own, but still assign her own views some weight, so that $C(A|q(A) = x)$ falls between x and $\frac{1}{2}C(A) + \frac{1}{2}x$." So even if one concedes that Nissan-Rozen's argument works in sufficiently non-ideal cases, it isn't clear that it would thereby refute a position that anyone subscribes to.

Nevertheless it's still possible appreciate the puzzle for the creative manner in which it challenges naive intuition (even in the case where Jane is taken to be ideally rational–an assumption we will make going forward). Nissan-Rozen writes of his paper that it "emerged from an exercise in Bayesian epistemology" he wrote for a course at the Hebrew University, and he thanked his students 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." We think that the puzzle serves this end well, and that we have a good line on the issues arising in it. Follows now our attempt to communicate it.

Nissan-Rozen entertains then rejects the idea that $c(T|HX)$ should be low in the following passage:

> 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.

As noted, denying lowness of $c(T|HX)$ when x is low requires one to renounce *Epistemic Deference*, a radical and plainly unwarranted move in the case we are interested in (where Jane is ideally rational). Still, the intuition that $c(T|HX)$ ought to be high is stubborn, even in this case. It's *wrong* of course–but how can one (more perspicuously than does the independence argument) defuse it?

We're sure there are many ways, but we think we've found an interesting one. Namely, via a "rarity of defeat" principle (of independent interest) that, to our minds, wholly and cleanly undercuts the stubborn intuition. The principle will be established using two features (the first entails the second) that apply whenever an agent C with credence function c defers to an expert J having credence function j and A is a measurable event. Denoting C's expectation operator by E, they are:

(i) $E(1_A|j(A) \leq x) \leq x$; and
(ii) $E(j(A)) = c(A)$.

Here now is our formulation of the principle.

Theorem 1. (Rarity of Defeat) Let A be a measurable event. Then for any $x < 1$ with $0 < x \leq c(A)$,

$$c(j(A) \leq x|A) \leq \frac{x(1 - c(A))}{(1 - x)c(A)}. \tag{1}$$

Proof. Denote by $\mathbf{X}$ the event $j(x) \leq x$. (ii) says that

$$c(A) = E(j(A)) \leq xc(j(x) \leq x) + 1 \cdot c(j(x) > x),$$

i.e. $c(A) \leq xc(\mathbf{X}) + 1 - c(\mathbf{X})$, from which it follows that

$$c(\mathbf{X}) \leq \frac{1 - c(A)}{1 - x}. \tag{2}$$

(i) meanwhile says that $c(A|\mathbf{X}) \leq x$. Together with (2) this yields

$$c(A\mathbf{X}) \leq xc(\mathbf{X}) \leq \frac{x(1 - c(A))}{1 - x}.$$

Dividing both sides of this equation by $c(A)$ gives (1). qed

Theorem 1's estimate is sharp; if J has conditioned on the partition

$$\left\{A \wedge \left(y \leq \frac{x(1 - c(A))}{(1 - x)c(A)}\right), \neg A \vee \left(y > \frac{x(1 - c(A))}{(1 - x)c(A)}\right)\right\},$$

where y is independently and uniformly distributed on $(0, 1)$, there is equality in (1).

The rarity of defeat principle gets its name from cases where $c(A)$ is near 1 and one seeks lower bounds on the confidence one should have that one *knows* that A. Suppose an ideal C's current credence in A is .999, and we subscribe to the formalism that KA if and only if A is true and C's credence in A will justifiably reach 1 eventually while never falling below .9. Taking J to be the future time slice of C where her credence in A is at its global future minimum (assuming that J's credence function j always has the form $j(\cdot) = c(\cdot|\mathcal{P})$ for a measurable partition $\mathcal{P}$), Theorem 1 says:

$$c(\neg KA) = c(\neg A) + c(A)c(j(A) \leq .9|A) \leq .001 + (.999)\frac{.9(1 - .999)}{(1 - .9).999} = .01.$$

Therefore C has at least 99% confidence that she knows that A.

The point at which the argument we are critiquing falls apart in the case where Jane is ideally rational can be isolated from the following excerpts:

1. "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)..."
2. "Jane might believe H to a low degree for many reasons. Jane might know that John lives in London or..."

Taken by itself, 1. is a clear non-sequitur; all that follows from the parenthetical observation is that $c(T|H)$ is very high. 2. looks like

an attempt to explain this away by minimizing the impact of having observed X. We can check to see if this works when x is low by letting $\mathbf{X}$ denote the event $j(H) \leq x = 2c(H|\neg T)$ and writing:

$$\begin{aligned} c(T|H\mathbf{X}) &= c(T|H)\Big(\frac{c(\mathbf{X}|TH)}{c(\mathbf{X}|H)}\Big) \\ &= c(T|H)\Big(\frac{c(\mathbf{X}|TH)}{c(\neg T|H)c(\mathbf{X}|\neg TH) + c(T|H)c(\mathbf{X}|TH)}\Big). \end{aligned}$$

By the intuitions we are scrutinizing this too is "very high". Of course we concede that $c(T|H)$ is very high, so the second summand in the denominator of the rightmost expression is essentially equal to the numerator. So this reasoning requires that the first summand should be negligible relative to the second. In particular, it requires:

Key Premise: $c(\neg T|H)c(\mathbf{X}|\neg TH)$ is negligible relative to $c(\mathbf{X}|TH)$.

We doubt that anyone would find the key premise appealing because they think that $c(\mathbf{X}|\neg TH)$ is small. It shouldn't be that surprising that Jane would have credence in H less than or equal to twice $c(H|\neg T)$ conditional on $\neg TH$, because in such a case $\neg T$ would be part of her evidence. (She could have additional evidence raising her credence in H above $2c(H|\neg T)$, but not necessarily.) Rather, we think that what gives the above premise its first blush appeal is the intuition that $c(\neg T|H)$ should be negligible relative to $c(\mathbf{X}|TH)$. Indeed, we think that is the point excerpt 2. is getting at. "Jane might believe H to a low degree for many reasons" looks to be an imprecise way of conveying that we should not be that surprised should Jane turn out to have such a low degree of belief, even in a case where both T and H are true.

According to the rarity of defeat principle, however, $c(\neg T|H)$ *can't* be negligible relative $c(\mathbf{X}|TH)$ if Jane is an ideal rational agent. This is because Jane's credence function j is expert with respect to $k(\cdot) = c(\cdot|T)$, so Theorem 1 says that

$$\begin{aligned} c(\mathbf{X}|TH) = k\Big(j(H) \leq x|H\Big) \;&\leq\; \frac{x(1-k(H))}{(1-x)k(H)} \\ &= \frac{2c(H|\neg T)(1-c(H|T))}{(1-2c(H|\neg T))c(H|T)} \\ &= \Big(\frac{2c(\neg TH)(c(T)-c(TH))}{c(H)c(\neg T)}\Big)\Big(\frac{c(H)}{(1-2c(H|\neg T))c(TH)}\Big) \end{aligned}$$

The second factor in the final expression is ≈ 1, and presumably $c(T) \leq \frac{1}{4}$ (this is generous; we know essentially nothing about John, and far

less than a quarter of people in the population have published papers in physics journals), so

$$c(\mathbf{X}|TH) < 3c(\neg T|H)\frac{c(T)}{c(\neg T)} \leq c(\neg T|H).$$

So it isn't just that $c(\neg T|H)$ isn't negligible relative to $c(\mathbf{X}|TH)$...it's in fact *always larger* than $c(\mathbf{X}|TH)$. Indeed, in most scenarios it's probably *much* larger, because even near equality holds in Theorem 1 only in very special circumstances.

This analysis also shows how proneness to irrationally extreme credences on Jane's part might make room for a scenario in which $c(T|H\mathbf{X})$ is in fact high. For what Theorem 1 shows to be rare is "justified defeat" or "defeat by the evidence"; defeat by one's own irrationality needn't be rare at all. So, if it's not that unlikely that Jane's credences in H have been decimated by evidence that is less supportive of $\neg H$ than she believes (John lives in Rhode Island, say, or spends a fair number of weekend hours playing soccer), the key premise could turn out to be true. Under such an assumption, one would, upon learning that T, H and X were all true, come to believe it likely that Jane had adopted a gratuitously low credence in H.

References

Elga, A. 2007. Reflection and disagreement. *Nous* 41:478-502.

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.

Nissan-Rozen, I. 2018. A puzzle about experts, evidential screening-off and conditionalization. *Episteme*. Forthcoming.