Testimony as Evidence

Katie Steele

May 2, 2011

Abstract

This paper considers a special case of belief updating—when an agent learns testimonial data, or in other words, the beliefs of others on some issue. The interest in this case is twofold: 1) averaging methods (linear/geometric) for updating on testimony are somewhat popular in epistemology circles, and it is important to assess their normative acceptability, and 2) this facilitates a more general investigation of what it means for any updating method to have a suitable *Bayesian* representation (taken here as the normative standard). The paper initially defends averaging methods against Bayesian-compatibility concerns raised by Bradley (2007). We further suggest that averaging be coupled with Jeffrey conditioning, if the aim is a fully comprehensive belief-update rule. This move, however, reveals important further questions: Is learning testimony tantamount to *changing one's mind*, or should it be seen as just another piece of *evidence* that contributes to an agent's beliefs? We side with the latter. A class of updating methods is identified that *does* treat testimony as evidence, but this class excludes averaging.

1 Introduction

This paper has two goals. The first, more specific goal is to investigate how an agent may update their beliefs in response to a particular kind of information—the testimony of others. Testimony is here understood as another agent's (Bayesian) beliefs, whether the beliefs concern direct experience, or a more involved inference, for instance, that there is life elsewhere in the universe. The second goal of the paper is more general: to investigate what it means for a belief-updating method to be compatible with the Bayesian model, and furthermore, what *type* of Bayesian learning the method expresses. The testimony case is useful for investigating these general issues, because testimony has been treated as a special case of learning in the literature, and has inspired alternative, not-obviously-Bayesian, belief-updating methods.

A few words on the relevant testimony literature is in order. One point of connection is the recent debate concerning how to resolve 'peer disagreement' that persists after all evidence has been shared (see, for instance, Feldman 2004, Elga 2007, Christenson 2007 and Kelly forthcoming). There is also a more formal literature (the actual starting point for this paper) concerning models of learning that specifically handle input data in the form of others' probabilistic beliefs at a time (see Lehrer and Wagner 1981, French 1985, Genest and Zidek 1986, Clemen and Winkler 1999).[1] In this literature too, it is assumed that agents have, as far as possible, shared all relevant data—at the time in question, they are in *reflective equilibrium*, to use the terminology of Lehrer and Wagner (1981). We will follow suit and restrict attention to cases where agents have shared all background evidence, at least to the best of their knowledge. Of course, the implicit assumption is that the agents did not have the same prior beliefs over an entirely identical probability space, and/or the same interpretation of the shared evidence, otherwise there would be no persistent disagreement.[2] Despite these differences, it is plausible that agents may yet regard each other as epistemic peers, in the sense that they *learn* from each others' beliefs.[3]

One might be concerned about just what counts as testimony, and whether it is really distinct from other kinds of data. Surely there are multitudes of more and less explicit ways to learn of another's beliefs, such as observing their behaviour, or mere traces of their behaviour. For example, the fact that my friend has taken her umbrella tells me something about her belief that it will rain. Should all these be treated as special cases of learning testimony? Perhaps the best response to this question is: yes. Even if it is gleaned in a variety of ways, testimony can be similarly represented as a statement of the beliefs of others. That is, testimony is not 'raw' data, but is the result of an implicit initial inference, as shown below:

event (e.g. communication) $\rightarrow$ initial inference $\rightarrow$ testimony

Our testimony problem is then as follows: the agent in question has a prior probabilistic belief function over a sigma algebra of events. The agent

[1]The cited models actually play a more ambitious role; they are intended to model the process of *consensus*, where a number of group members update on each others' beliefs and arrive at the same belief. Here we are concerned only with the proposed methods of updating, and not the conditions under which consensus is achieved within a group.

[2]Likewise, there can be no ultimate disagreement if agents have the same prior probability function and have common knowledge of each other's posteriors, according to Aumann's (1976) theorem.

[3]This is a slightly more general interpretation of 'epistemic peer' than in the informal peer disagreement debate; in that debate, an epistemic peer is often taken to be someone whose beliefs one respects *equally* to one's own.

receives some testimonial data *at some point*, i.e. they learn the beliefs, regarding the partition $\mathbf{B}$, of n other agents, the respective belief functions being P_i, $i = 1 \dots n$. (These others need not entertain the same full probability space as our agent of interest; they must merely have beliefs on $\mathbf{B}$.) By way of representation, we can construct a matrix $M_{\mathbf{B}}$ with columns corresponding to the events in $\mathbf{B}$ and rows corresponding to the belief functions of the expert peers. (We can refer to this matrix as the testimony profile.) The task is to specify the principle agent's posterior belief function. That is, what should be the updating function $F_{\mathbf{B}}$ for determining our agent's new beliefs across $\mathbf{B}$, given their prior beliefs and the testimony profile $M_{\mathbf{B}}$? To state the problem semi-formally, we seek plausible candidates for:

$F_{\mathbf{B}}$: prior, $M_{\mathbf{B}} \rightarrow$ posterior

Recall our assumption that the agents have shared evidence, to the best of their knowledge. Note also that our focus is restricted to a one-off interaction with any particular group of peers—the agent has not already received testimony from these peers sometime in the past. This limitation will be revisited briefly in the conclusion.

The initial aim in what follows is to examine the functions $F_{\mathbf{B}}$ that have become popular in both the peer disagreement and formal consensus literature—weighted linear and geometric averaging. (To be more precise, the focus here is on linear averaging, but much of the discussion applies to geometric averaging as well.)[4] Not only has testimonial data been taken as a distinct form of data, it has also been considered worthy of special treatment vis-à-vis updating. Section 2 briefly comments on why this might be so. The rest of the paper examines whether and how linear averaging may be compatible with Bayesian updating. Section 3 takes a preliminary stance on compatibility, and argues against Bradley (2007) that an instance of linear averaging *can* be considered prima facie compatible with an instance of Bayesian conditioning. There are some lingering concerns about the versatility and completeness of the averaging method, that are addressed in Section 4, by appeal to a two-step procedure involving Jeffrey-conditionalization. This move, however, leaves open some important questions. Section 4 argues that learning testimony is not a matter of *changing one's mind*; testimony should be seen, rather, as just another piece of *evidence* that contributes to an agent's beliefs. A class of updating methods is identified in Section 5 that *does* treat testimony as evidence, but this class excludes averaging.

[4]Weighted averages are prominent in the formal consensus modelling literature. Moreover, a crude kind of average is arguably the favoured solution in the peer disagreement debate—what is referred to as the 'equal weights' view.

2 Why not Bayesian business as usual?

Before turning to an examination of linear averaging for updating on testimony, it is helpful to consider why alternatives to Bayesian conditionalization might have been proposed in the first place. The aim of this section is merely to offer some initial motivation for departures from the standard Bayesian model.

The Bayesian model holds that the pooling function $F_{\mathbf{B}}$ should accord with Bayes' formula:

$$P'_0\left(B_j \mid P_1(B_j) = p_1, P_2(B_j) = p_2, \ldots, P_n(B_j) = p_n\right)$$

$$= \frac{P_0(P_1(B_j)=p_1, P_2(B_j)=p_2, \ldots, P_n(B_j)=p_n \mid B_j) \times P_0(B_j)}{P_0(P_1(B_j)=p_1, P_2(B_j)=p_2, \ldots, P_n(B_j)=p_n)} \qquad \forall B_j \text{ in } \mathbf{B}.$$

where
$P_0(B_j)$ and $P'_0(B_j)$ are the agent's prior and posterior for event B_j
and $P_i(B_j)$ is witness i's probability for event B_j (at the time in question)

The 'problem' with this function, or the reason why some might consider it not sufficiently user-friendly, is that the testimony of others is not combined directly with the agent's own probabilistic beliefs; belief change is governed, rather, by the relevant likelihoods

$$P_0\left(P_1(B_j) = p_1, P_2(B_j) = p_2, \ldots, P_n(B_j) = p_n \mid B_j\right)$$

which represent the agent's belief that their peers would have the beliefs specified, conditional on the truth of each event B_j in $\mathbf{B}$. (The likelihoods conditional on the falsity of each B_j also play a role, of course.) We do not here deny that Bayesian conditioning is the most accurate way to represent rational belief change in response to testimony, in a sufficiently detailed model; indeed the Bayesian model is treated as the normative standard in this paper. The point is just that the Bayesian model may be somewhat cumbersome (with respect to the number of propositions that must be modelled) and also awkward to use. Indeed, the Bayesian expression above treats testimony just like any other type of evidence—an event that is merely indicative of the truth/falsity of the events B_j under consideration.

Averaging methods may have become popular 'shortcuts' for updating on testimony, precisely because, in contrast to the above, the probabilistic beliefs of others play a direct role in these functions. We see this by considering the formal statement of the weighted linear average:

$$P'_0(B_j) = w_0 \times P_0(B_j) + w_1 \times P_1(B_j) + \ldots + w_n \times P_n(B_j)$$

where $w_0, \ldots, w_n$ are interpreted as the 'weights of respect' assigned to all agents involved, and are non-negative and summing to one.

The posterior belief on $\mathbf{B}$ for the agent in question is a linear 'pool' of the actual beliefs of all agents involved. Instead of treating other agents' beliefs like a litmus test for determining the truth of the events B_j, the agent takes these beliefs on board directly; the agent mixes these beliefs directly with their own. Weighted linear/geometric averaging allows this to be done in such a way that the beliefs of those the agent most respects have the most influence, or are most dominant in the mix. This is, *prima facie*, a more natural or user-friendly way to respond to the beliefs of others.[5]

The popular defence for using linear averaging, in particular, to serve as the shortcut function for updating on testimony ($F_{\mathbf{B}}$), is given mathematically in Wagner (1985). Lehrer and Wagner (1981) mount the same defence more explicitly in the context of the updating problem as opposed to the group aggregation problem. In short, linear averaging is the only function to satisfy the *Independence of Irrelevant Alternatives* condition, (*IA*). The appeal to IA amounts to a *multi-profile* justification: if we consider all possible combinations of probability functions for the agents involved, IA states that where the vector of probabilities for a single event B_j are equivalent, the posterior belief for B_j should be equivalent. In other words, the posterior for a single event B_j depends just on the probabilities all agents assign to B_j, and it does not matter what are the complete belief functions of these agents. Wagner (1985) proves that if IA alone is stipulated (under *universal domain*), then $F_{\mathbf{B}}$ must be a weighted linear average, with some error term. The function is restricted to positive weights and zero error (as per the expression above) if the further condition of *Zero Preservation*, (*ZP*), is stipulated; ZP states that if all agents assign probability zero to some event B_j, then the posterior for B_j should also be zero.

One might regard the above two-step explanation a bit quick, i.e. that there is sufficient motivation for a special method for updating on testimony (which is moreover a distinct type of information), and that this method should be a weighted linear average, as it must satisfy IA and ZP. Indeed, fans of the weighted geometric average would reject the latter desiderata.[6] Even if one sees little positive motivation for linear averaging, however, the fact remains that the method has a significant presence in the literature,

[5]Like others, (e.g. Lehrer and Wagner 1981), we assume that an agent may 'respect' another's beliefs at some point in time, despite not necessarily having the same prior, nor the same interpretation of the shared evidence.

[6]Indeed there is some debate about the comparative merits of linear and geometric averaging; see Genest and Zidek 1986, Clemen and Winkler 1999, Shogenji 2007.

and it is important to consider whether this updating approach is at least rationally *permissible*. To this end, the next sections examine whether linear averaging is compatible with the Bayesian model.

3 The initial compatibility challenge

The task of determining whether linear averaging is compatible with Bayesian conditionalization mostly involves working out what it *means* for the two updating methods to be compatible. Our initial examination of compatibility in this section is confined to a single testimony update in isolation.[7] To begin with, we note that linear averaging and conditionalization are not isomorphic. Linear averaging is prima facie a special case of Bayesian conditionalization—any posterior obtained via an instance of linear averaging can also be obtained via *some* Bayesian model. This is because, given a prior probability function on the partition $\mathbf{B}$, *any* posterior probabilities on $\mathbf{B}$ may be obtained, if we may choose the likelihoods on $M_{\mathbf{B}}$, where $M_{\mathbf{B}}$ is the testimony. Consider:

$$\frac{P_0'(B_j)}{P_0' \neg B_j} = \frac{P_0(M_{\mathbf{B}}|B_j)}{P_0(M_{\mathbf{B}}|\neg B_j)} \times \frac{P_0(B_j)}{P_0(\neg B_j)}$$

If the likelihood ratio for each B_j may vary, and thus also the unconditional probability of $M_{\mathbf{B}}$, then any transition from the ratio of priors to the ratio of posteriors is possible, unless the priors take extreme values of zero or one.

The converse does not hold, i.e. not all Bayesian updates on testimony can be expressed as linear averages over the relevant partition. For example, there is some Bayesian model, yet no linear averaging model, that permits an agent the following update:

prior $[\ 1/6 \quad 1/3 \quad 1/2\ ]$ plus testimony $[\ 1/2 \quad 1/6 \quad 1/3\ ] \longrightarrow [\ 1/3 \quad 1/3 \quad 1/3\ ]$

The fact that linear averaging is restrictive is not necessarily a bad thing, however, provided an explanation can be given as to when averaging is applicable, and why scenarios like that above do not arise in such cases. For

[7] Our discussion here of Bayesian compatibility thus differs from those of Shogenji (2007) and Jehle and Fitelson (2009). These authors consider whether linear averaging satisfies the so-called 'Conditionalization' criterion (or what Wagner calls 'Bayesianity'). Their Conditionalization criterion does not concern a single testimony update in isolation, but rather the relationship between testimonial data and other shared evidence—whether it matters if the agent updates on testimony before or after all parties receive some evidence in their common proposition space. Note that linear averaging does not satisfy this criterion, while geometric averaging does.

instance, one stipulation is that the agent must regard those giving testimony as epistemic peers, in the limited sense that they would assign them positive respect.

A more significant worry regarding compatibility would be the lack of a one-to-one map from particular linear average updates to Bayesian updates. Bradley (2007) effectively argues that there *is* such a mapping problem for linear averaging. This section defends averaging against Bradley's and similar criticisms, by showing that the problem can be dodged by attending carefully to the universal domain condition that grounds the averaging method. There remain further problems for linear averaging, but these must wait for the next section.

A number of authors question how 'weights of respect' in linear averaging should be interpreted and ascertained (e.g. French 1985), but Bradley (2007) expresses a more fundamental worry about these weights: If we appeal to the multi-profile justification of linear averaging given in the last section, then the respect weights must be constant for group members regardless of the nature/origins of the beliefs they express, and thus the method ignores distinctions that are important in the Bayesian model. That is, one linear average model maps to more than one Bayesian model, and the differences between the latter set are important.

In particular, Bradley argues that averaging treats independent agents identically to dependent agents, when the Bayesian model treats them differently. Let us rehearse the argument. Following Bradley, we will keep things simple and assume there are just two agents giving testimony to the principle agent. The latter learns the probabilities for these two agents across partition **B**. According to the Bayesian model, our agent's new probability for one event in **B**, call it b, given this new information, is:

$$P_0'\,(b \mid P_1(b) = p_1,\ P_2(b) = p_2) \;=\; \frac{P_0(P_1(b){=}p_1,\, P_2(b){=}p_2|b)\times P_0(b)}{P_0(P_1(b_1){=}p_1,\, P_2(b){=}p_2)}$$

If the probability functions for the two consulted experts are independent given b, and are also unconditionally independent, the above equals:

$$= \frac{P_0(P_1(b){=}p_1|b)\times P_0(P_2(b){=}p_2|b)\times P_0(b)}{P_0(P_1(b){=}p_1)\times P_0(P_2(b){=}p_2)}$$

At the other extreme, if the probability functions for the two experts are perfectly correlated, the expression equals:

$$= \frac{P_0(P_1(b){=}p_1|b)\times P_0(b)}{P_0(P_1(b){=}p_1)}$$

Bradley notes that these two expressions will only be equal if

$\frac{P_0(P_2(b)=p_2|b)}{P_0(P_2(b)=p_2)} = 1$

which is to say that the principle agent believes one of the agents' beliefs to be independent not only of the other agent, but also independent of the truth. In that case, by any reasonable interpretation of weights of respect, this agent should be given a weight of zero in the linear averaging function.

We see from the above that if the domain of an averaging method is any probability profile for a group, including cases where the experts consulted have independent beliefs as well as cases where their beliefs are thought to be dependent in some way, then linear averaging will only be consistent with the Bayesian model in special circumstances—when the set of respect weights mirrors the situation where the beliefs of all consulted experts except one are independent of the truth (the trivial case as far as independence versus dependence amongst experts is concerned). This sort of restriction defeats the purpose of providing a model that allows an agent to update on the beliefs of *a number* of other agents who they regard as epistemic peers.

A response can be made to the criticism above, that does not involve sacrificing the multi-profile justification of linear averaging. The trick is to carefully specify the domain over which the IA condition must hold, and consequently the domain over which the weights of respect are constant. Consider the following situation: Our agent may consult Group 1, constituted by experts whose beliefs about $\mathbf{B}$ are independent, or else our agent may consult Group 2, constituted by experts whose beliefs have some pattern of dependency. The IA/universal-domain condition effectively requires the same respect weights be assigned to members of Group 1 (or 2), whatever the members' belief functions happen to be. It does not, however, require matching respect weights across two different groups who express the same set of probability functions, unless the testimony profiles associated with these groups are not distinguished, but are rather part of the same domain. And this need not be the case.

This point about the domain of a particular linear averaging function applies more broadly than the case of independent versus (partially) dependent peer groups. One might also be worried that the same group of experts may have wildly different expertise with respect to different issues, and yet constant weights of respect will not reflect this. For instance, we would not want to assign the same respect weightings for a certain group of peers regardless of whether we were asking them about average rainfall for the next wet season or whether unemployment will drop. Again, the mistake here is to think that belief profiles concerning different issues/propositions are part of the same domain over which the justifying condition for linear averaging, IA, applies. It must simply be stipulated that IA applies only to sets of beliefs concerning the *same* event space. That is, if the partition $\mathbf{B}$ is being assessed, the universal domain spans all possible (prior, testimony profile) pairs for the group that are constituted by probability functions on $\mathbf{B}$. If, on the other hand, a different partition is in question, say $\mathbf{C}$, then the universal domain would span a different set of (prior, testimony profile) pairs, this time probability functions on $\mathbf{C}$.[8]

Given the points just made, a slight change in the notation used above is in order. Recall the general representation of the testimony problem given in the first section:

$F_{\mathbf{B}}$: prior, $M_{\mathbf{B}} \rightarrow$ posterior

The use of the $\mathbf{B}$ index reflects the point above that the matrix of probability functions in question pertains to a specific issue or partition—the $\mathbf{B}$ issue. The updating function (in particular the respect weights) are specific to that issue, hence $F_{\mathbf{B}}$. But we might want to make explicit the first point as well, that the matrix is specific to a particular context—a group of peers at a particular point with a believed pattern of dependency in their beliefs. This rich context might be represented by the further index $G \in \mathbf{G}$. So the testimony received by an agent is represented $M_{\mathbf{B},G}$, and likewise, the updating function that we seek is $F_{\mathbf{B},G}$.[9] Of course, all this indexing highlights how removed testimony is from raw experience in our model; the evidential statement $M_{\mathbf{B},G}$ is laden with inferences about the 'group context' that are not explicitly modelled. Moreover, there is still a big question as to how the group context is translated into weights of respect. We will not explore these details, however, as they are not the prime focus here.

[8]This raises the further question of how weights of respect on the $\mathbf{B}$ and $\mathbf{C}$ partitions relate to each other. For starters, if the $\mathbf{C}$ partition is a refinement of the $\mathbf{B}$ partition, then surely the $\mathbf{C}$-weights should add appropriately to the $\mathbf{B}$-weights. For the more interesting cases where the $\mathbf{C}$ and $\mathbf{B}$ partitions are orthogonal, and to some extent dependent, whether by the principal agent's reckoning or by their peers' reckoning, it gets more tricky. We will not elaborate further just now on relations between respect weights, however, because the prospect of a complex event space raises some bigger issues, as will become clear in the next section.

[9]Note that one could potentially appeal to a change in 'group context' as a way to save linear averaging from violations of the Conditionalization/ Bayesianity criterion that was mentioned in a previous footnote. The argument would be that the group context differs before and after all agents receive new evidence, and perhaps the respect weight vectors could be synchronised in order to satisfy the aforementioned criterion.

4 Rich event spaces and a new compatibility challenge

The reference above to different issues or partitions, say, **B** and **C**, raises the further question of how linear averaging is supposed to work in a rich algebraic setting. Unless this is clarified, averaging belief-update methods are at best incomplete. Arguably the most natural solution is that averaging be combined with Jeffrey-updating, such that, in response to testimonial data regarding **B**, probabilities are updated across this partition in line with linear averaging, and then the probabilities of all other propositions are subsequently updated so that probabilities conditional on the individual events of **B** remain constant or 'rigid'.

Note that the appeal to Jeffrey-conditionalization is a change of tack from the previous section, as we now depict averaging as an *extra*-Bayesian process. While the points about group context remain important, learning testimony is no longer likened to gaining knowledge of a single proposition E, as per strict Bayesian conditionalization.

It is best to illustrate with an example. Consider a simple setting where our agent has the following prior probability function $P_{0,\mathbf{D}}$ over the event space $\mathbf{D} = \mathbf{B} \times \mathbf{C}$:

	B	$\neg B$
C	0.1	0.2
$\neg C$	0.3	0.4

The agent meets an expert on the **B** partition, who has $P_{1,\mathbf{B}} = [0.9, 0.1]$. Linear pooling with, for example, weight 1 to this expert, gives $[0.9, 0.1]$ over the **B** partition. Clearly, however, this is not a complete specification of the agent's posterior probability function. This is where Jeffrey conditionalization enters. Accordingly, the agent's new probabilities over the entire space are:

	B	$\neg B$
C	0.225	1/30
$\neg C$	0.675	2/30

The two-step procedure—averaging then Jeffrey-conditionalization—can thus be regarded a fully comprehensive belief-update rule!

The problem of rich event spaces does not, however, end here. The well-known puzzles with Jeffrey-conditionalization—that it is not generally commutative with respect to changes in probabilities on different partitions—prompt further investigation of the proper treatment of testimony. Our earlier example can be extended to illustrate the non-commutativity property. Recall that our agent has already updated her beliefs with respect to **B** in response to the beliefs of an expert peer on **B**. After this encounter, assume that the agent meets a (different) expert on the **C** partition who has $P_{2,\mathbf{C}} = [0.6, 0.4]$. The obvious strategy is to apply the linear-plus-Jeffrey method a second time around. Linear pooling, again with the assumption of weight 1 for the **C**-expert, gives $[0.6, 0.4]$ over the **C** partition. Applying Jeffrey conditionalization, we get the following posterior:

	B	$\neg B$
C	0.52258	0.077419
$\neg C$	0.36404	0.035955

But what if the experts had made their reports in reverse order? In this case, we would get the following transition of probability functions (assuming the same weightings, i.e. 1 to each expert):

	B	$\neg B$			B	$\neg B$			B	$\neg B$
C	0.1	0.2	$\rightarrow$	C	0.2	0.4	$\rightarrow$	C	0.48461	0.063637
$\neg C$	0.3	0.4		$\neg C$	0.17143	0.22857		$\neg C$	0.41539	0.036363

Clearly the averaging-plus-Jeffrey update procedure may be sensitive to the order in which testimony is received.[10]

The question is whether non-commutativity is a bad thing in this context. Should updating on testimony from different sources, and pertaining to different partitions, combine in any order to yield the same posterior belief function? Or might updating on one partition partially override any previous updating on other partitions, as if the agent were simply changing her mind? The former, of course, is an appeal to commutativity, while the latter is not.

The work of Wagner (2002), which builds on Field (1978), Diaconis and Zabell (1982) and Jeffrey (1988), allows us to approach the question from a different angle. Wagner proves the following are *sufficient* conditions for changes in probabilities across two partitions to be commutative:

Consider the following two series of probability functions due to Jeffrey-updates across the partitions **B** and **C**:

[10] Order is not important, whatever the probability updates across each partition, just in case the partitions in question are probabilistically independent. Diaconis and Zabell (1982) show, moreover, that order may not be important for *some* probability updates across different partitions, even if the partitions are not probabilistically independent. They refer to this as *Jeffrey independence*—the label applies to particular partitions and particular probability updates across these partitions.

$P \to_{\mathbf{B}} Q \to_{\mathbf{C}} R$
$P \to_{\mathbf{C}} S \to_{\mathbf{B}} T$

The posterior probability functions R and T are identical if

$$\beta_P^Q(B_i, B_j) = \beta_S^T(B_i, B_j) \qquad \forall B_i, B_j$$

and

$$\beta_Q^R(C_k, C_l) = \beta_P^S(C_k, C_l) \qquad \forall C_k, C_l$$

where

$\beta_P^Q(A, B) = \frac{Q(A)}{Q(B)} / \frac{P(A)}{P(B)}$ $\qquad$ (P being the prior, Q the posterior, and β the 'Bayes factor')

Wagner uses the conditions for commutativity to inform a definition of *identical* Jeffrey-style learning—two cases of learning are identical if they are characterized by the same set of Bayes factors, as defined above. On this reading, a change on the $\mathbf{B}$ partition to $[b_1, b_2]$ followed by a change on the $\mathbf{C}$ partition to $[c_1, c_2]$ does *not* generally amount to the same sequence of learning (only in reverse order) as a change on $\mathbf{C}$ to $[c_1, c_2]$ followed by a change on $\mathbf{B}$ to $[b_1, b_2]$. This is because the sets of relevant probability changes on the partitions, as described by the Bayes factors, may differ.[11]

While Wagner may be interpreted as providing a purely formal definition of identical learning, Field (1978) makes the more substantial point, with respect to Jeffrey-style learning, that the same sensory input should in fact result in identical learning—it should have a particular *evidential impact* on an agent's priors, as characterized by the set of Bayes factors. Unfortunately, however, this nice simple relationship between sensory experience and Bayes factors turns out to be *too* simple, as shown by Garber (1980); in short, it does not take account of dependencies amongst these experiences. In the context of strict conditionalization, learning E_1 followed by E_2 yields the same posterior as learning E_2 followed by E_1, but the sets of Bayes factors may differ if E_1 and E_2 are not independent, given the propositions of interest. Likewise, in the context of Jeffrey conditionalization, we cannot simply assume that a particular sensory experience will have a constant impact, in terms of Bayes factors, on the prior probability function. Moreover, there is the more basic dispute about whether Jeffrey-style learning must even be

[11]Refer back to previous footnote on probabilistic independence and Jeffrey independence.

commutative.[12]

It may be that Jeffrey-updating is flexible—it can accommodate instances whereby the agent simply changes her mind on some issue, and it can also accommodate incremental learning, where the agent's beliefs are based on her full stock of sensory experience. Surely testimony, however, fits the latter mold. (Of course, we noted in the introduction that testimony is not raw sensory experience in our model. That is not important. There is still the question of whether we want to treat updating on testimony as a case of changing one's mind, or as a case of responding to an extra piece of evidence.) When the agent receives the testimony of a group of peers, she learns something new, represented fully as $M_{\mathbf{B},G_i}$ (recall that G_i specifies the 'group context'). Surely this information should not simply be 'overridden' by any subsequent testimony report, such as $M_{\mathbf{C},G_j}$. This is to say that updating on testimony should be commutative. Updating on $M_{\mathbf{B},G_i}$ and then $M_{\mathbf{C},G_j}$, should yield the same posterior as updating on $M_{\mathbf{C},G_j}$ and then $M_{\mathbf{B},G_i}$. The averaging-plus-Jeffrey method is therefore problematic, because, as we have seen, it is not generally commutative.

It is less clear that a particular piece of testimony, say $M_{\mathbf{B},G_i}$, should always produce identical learning, à la Wagner, and thus have the same *impact* on an agent's priors. One can require commutativity without requiring constant Bayes factors for a particular 'sensory' experience, as noted above. Indeed, one might argue that when there are probabilistic dependencies between partitions, order really should matter to the relevant Bayes factors, even if order does not matter to the final posterior. The problem is that it is difficult to model commutativity for testimony, without stipulating that the one piece of testimony should be associated with the one set of Bayes factors. Indeed, we make this stipulation in the next section, while recognising that it may well be too restrictive.

5 Testimony as evidence

The previous section suggests a desideratum for an updating method to be compatible with Jeffrey conditionalization, and moreover treat the input as incremental evidence: commutativity. As forewarned, we accommodate this desideratum by imposing a stronger assumption, namely: same testimony input, same learning or evidential impact. In other words, the agent's shift

[12] Jeffrey's own examples for motivating his generalized rule of conditionalization are apparently intended to illustrate that commutativity is not necessary: the agent's experience of the cloth's colour in candlelight, for instance, simply 'overrides' her previous experience of the cloth's colour in starlight, and vice versa. One could say that the agent here does not gather evidence incrementally about the colour of the cloth; she simply changes her mind.

from prior to posterior over the relevant partition should yield the same Bayes factors for the same testimonial input.

Recall the general form for testimony updating functions:

$F_{\mathbf{B},G}$: prior, $M_{\mathbf{B},G} \rightarrow$ posterior

Our new criterion is:

$\frac{P'_0(B_i)}{P'_0(B_j)} / \frac{P_0(B_i)}{P_0(B_j)} = c_{i,j} \qquad \forall i, j$

That is, for any two events in the **B** partition, the updating function, given a particular testimony profile $M_{\mathbf{B},G}$, should be such that the ratio of posteriors for the events divided by the ratio of priors is a constant.

The above criterion is satisfied by testimony updating functions that take the form:

$P'_{0,\mathbf{B}} =$ normalize $[P_{0,\mathbf{B}} \times f(M_{\mathbf{B},G})]$

That is, the agent's prior probability over **B** is multiplied by some function of the belief profile for the group of witnesses, and then normalized.[13] (The subsequent step is Jeffrey-conditionalization.) Of note is that neither linear nor geometric averaging fit this functional form; both are ruled out by our desideratum.

Rather than seeking further desiderata to pinpoint a particular function for updating on testimony-as-evidence, it is more helpful to highlight a property common to all functions that have the form stated above. To this end, let us first introduce the term *defer to testimony*; we define it here as 'changing one's beliefs to match the *aggregate* testimonial input'. The aggregate testimonial input is given by the chosen function $f(M_{\mathbf{B},G})$. An agent cannot generally defer to testimony, in this sense, if they update in a manner that treats the testimony as evidence. Indeed, only when the agent's prior distribution on the partition in question, say **B**, is the 'flat' distribution (i.e. equal probabilities for all events in **B**), will updating on $M_{\mathbf{B},G}$ amount to deferring to this testimonial input (in accord with the function f). One could say: only when the agent is maximally uncertain with respect to some partition, do they defer to testimonial input.

To illustrate the above point, consider the following example update rule:

[13] Presumably this functional form is both necessary and sufficient for satisfying the desideratum, but whether it is necessary is not so obvious, and not explored here.

$$P'_{0,\mathbf{B}} = \text{ normalize } [P_{0,\mathbf{B}} \times f(M_{\mathbf{B},G})] = \text{ normalize } [P_{0,\mathbf{B}} \times \sum_{i=1}^{n} w_i \times P_{i,\mathbf{B}}]$$

The posterior is not the linear average of all probability functions on **B**, including the agent's prior; here it equals, rather, a weighted average of the testimonial probabilities alone, *multiplied* by the agent's prior. When the prior on **B** represents maximal uncertainty, i.e. the flat distribution, then the agent defers to testimony: in this case their posterior is equal to the weighted average of the testimonial evidence.

Two example cases allow a clearer picture of the updating rule just specified vis-á-vis deferral:

Case 1

The agent has maximally uncertain beliefs regarding a two-event partition, i.e. their probability distribution is $[0.5, 0.5]$. They receive the following testimony: two experts to whom they give equal respect have probability distributions $[0.1, 0.9]$ and $[0.3, 0.7]$ respectively. Because the agent started with a flat prior, they effectively *defer* to the testimony: the agent's posterior will just be the equally weighted average of the two expert probability distributions, i.e. $[0.2, 0.8]$.

Case 2

The agent has some prior opinion with respect to a 2-event partition: their probability distribution is $[0.2, 0.8]$. The agent receives testimony from a peer for whom they have *maximal* respect, i.e. a weighting of 1. The peer has a probability distribution on the relevant partition of $[0.1, 0.9]$. Even though the agent has maximal respect for this peer, they do not defer to them. The agent's new probability function is rather: *normalize*$([0.2 \times 0.1, 0.8 \times 0.9])$.

The latter example makes vivid that maximal respect/consideration for any given testimony does not, in most cases, translate to deferral, given an updating function that treats testimony as evidence. So if I were to meet Stephen Hawking (my physics expert, let's say) for the first time, and he told me that his probability for string theory was 0.7, then I could only permissibly change my belief for string theory to 0.7, in line with his belief, if I initially assigned equal probability to this theory and its competitors. This may seem counter-intuitive—surely my prior opinion on string theory no longer matters, in the face of Hawking's. But we might well ask: Why should my previous opinions *not* matter? Note that Hawking's views may yet have a large impact on my beliefs, by an updating function that treats his testimony as evidence. And we might apply some backward reasoning: if

I knew in advance that I was prepared to update my beliefs on string theory to whatever Hawking said, then this suggests my prior beliefs on the issue really should have taken the form of the flat distribution.[14]

6 Concluding remarks

We initially set out to assess the normative acceptability of linear averaging, with the broader aim of determining just what normative acceptability entails. The previous section specified a criterion for an *extra*-Bayesian rule for updating on testimony: testimony updates should be commutative, and to achieve this, the same testimony should have the same evidential impact on the agent's priors, à la Wagner.

This criterion is not satisfied by straight linear averages. We saw, however, that an average of sorts *can* be accommodated—an average of the testimonial probability functions alone, which would then be multiplied by the principal agent's prior. So the popular defence of averaging given in Section 2—that it is more 'user-friendly' than fully detailed Bayesian conditionalization, and that it satisfies an *Independence of Irrelevant Alternatives* criterion, as well as *Zero Preservation*—may yet be relevant. In that case, our response to Bradley in Section 3 would also remain important, not just for the general insights about 'group context', but also for the clarification of the IA criterion.

Of course, other updating functions also satisfy the criterion identified in Section 5; we have not here argued for one in particular. That may be a project for later. More importantly, a number of details have not been addressed in this analysis, in the interests of making a simple case for treating testimony as evidence. We have assumed a rather naïve set-up, where an agent meets some peers for the *first* time, has already engaged in some (perhaps informal) sharing of evidence, yet still finds there is disagreement on some issue, hence the updating on testimony. Sections 4 and 5 introduced updates across more than one partition (of the principle agent's probability space), but again, a simple approach was taken. There is an implicit background assumption that entirely different groups of peers provide testimony on the different partitions. The situation is more complex if the same peers provide testimony on more than one issue. Moreover, complications arise when the same peers provide testimony on the same issue again and again, whether or not new evidence has been received in the meantime. In all these cases, more attention must be given to individuating 'group contexts' and

[14] In general, if testimony is treated as evidence, then the requirement that one's present beliefs be the expectation of one's future beliefs applies here too.

determining any further constraints on the relationships between the testimonial updates. We leave these issues, however, for future investigation.[15]

References

Aumann, Robert J. 1976. Agreeing to Disagree. *The Annals of Statistics* 4(6):1236–1239.

Bradley, Richard. 2007. Reaching a Consensus. *Social Choice and Welfare* 29:609–632.

Christensen, David. 2007. Epistemology of Disagreement: the Good News. *The Philosophical Review* 116(2):187–217.

Clemen, Robert T. and Robert L. Winkler. 1999. Combining Probability Distributions From Experts in Risk Analysis. *Risk Analysis* 19(2):187–203.

Diaconis, Persi, and Sandy L. Zabell. 1982. Updating Subjective Probability. *Journal of the American Statistical Association* 77(380): 822–830.

Elga, Adam. 2007. Reflection and Disagreement. *Noûs* XLI(3):478–502.

Feldman, Richard. 2004. Reasonable Religious Disagreements. manuscript URL http://www.ling.rochester.edu/ feldman/papers/reasonablereligiousdisagreements.pdf

Field, Hartry. 1978. A Note on Jeffrey Conditionalization. *Philosophy of Science* 45(3): 361–367.

French, Simon. 1985. Group consensus probability distributions: A critical survey. J. M. Bernado, M. H. DeGroot, D. V. Lindley and A. F. M. Smith (eds.) *Bayesian Statistics Vol. II.* Amsterdam: North-Holland. 183–197.

Garber, Daniel. 1980. Field and Jeffrey Conditionalization. *Philosophy of Science* 47(1): 142–145.

Genest, Christian, and James V. Zidek. 1986. Combining probability distributions: a critique and an annotated bibliography. *Statistical Science* 1:114–148.

Jeffrey, Richard. 1988. Conditioning, Kinematics, and Exchangeability.

[15]Thanks to . . .

B. Skyrms and W. Harper (eds.) *Causation, Chance, and Credence* vol 1. Dordrecht: Kluwer. 221–255.

Jehle, David, and Branden Fitelson. 2009. What is the "Equal Weight View"? *Episteme* 6(3): 280–293.

Lehrer, Keith, and Carl Wagner. 1981. *Rational consensus in science and society.* Dordrecht: Reidel

Shogenji, Tomoji. 2007. A Conundrum in Bayesian Epistemology of Disagreement. unpublished manuscript.

Wagner, Carl. 1985. On the formal properties of weighted averaging as a method of aggregation. *Synthese* 62:97–108.

Wagner, Carl. 2002. Probability Kinematics and Commutativity. *Philosophy of Science* 69:266–278.