# An Accuracy Based Approach to Higher Order Evidence

Miriam Schoenfield - Draft of 12/1/14

References and acknowledgements incomplete

## 1. Introduction

HYPOXIA: Aisha is flying her airplane on a bright Monday morning, wondering whether she has enough gasoline to get to Hawaii. Upon looking at the dials, gauges and maps she obtains some first order evidence E, which she knows strongly supports (say to degree 0.99) either that she has enough gas (G) or that she does not have enough gas (~G). Aisha does some complex calculations and concludes G, which is, in fact, what E supports. But she then gains some *higher order evidence* (evidence that bears on her own rationality): she realizes that she's flying at an altitude that puts her at great risk for hypoxia, a condition that impairs one's reasoning capacities. Aisha knows that pilots who do the kind of reasoning that she just did, and who are flying at her current altitude, only reach the correct conclusion 50% of the time.[1]

How confident should Aisha be that she has enough gas? This is the sort of question that the higher order evidence debate has been concerned with and there are two positions in the debate that will be the primary focus of this paper.

*The calibrationist* thinks that Aisha's first order evidence (dials, gauges, maps) supports a .99 credence in G (as stipulated), but her *total* evidence supports a 0.5 credence in G. Even though Aisha, in fact, drew the correct conclusion from her first order evidence, she can't rationally be confident that she did. After all, she has good reason to think that she's hypoxic![2]

The *steadfaster* thinks that Aisha's higher order evidence is completely irrelevant to the question of whether G, and that her *total* evidence also supports a 0.99 credence in G. After all, what possible bearing could facts about Aisha's mental states have on the likelihood that there is enough gasoline in the tank?[3]

---

[1] This case is based on a case from Elga (ms.)

[2] Some calibrationists include Christensen (2010), Elga (ms.), and Horowitz and Sliwa (ms.).

[3] Some steadfasters include Kelly (2010), Lasonen Aarnio (2014) and Weatherson (ms.).

While the higher order evidence debaters have been developing a variety of interesting arguments for their respective positions[4] there has been an increasing interest in what is sometimes called *accuracy-first* epistemology: the project of deriving rational requirements from accuracy based considerations.[5] The goal of this paper is to apply the accuracy-based approach to the case of higher order evidence.

Before getting into the technical details, let's think intuitively about which view of higher order evidence an interest in accuracy might motivate. For now, we can characterize an interest in accuracy as an interest in having high credences in truths and low credences in falsehoods.

Here's the first thing to note: Intuitively, it seems reasonable for us to expect, even a priori, that if Aisha responds to her evidence as the calibrationist recommends, she will be less accurate than she would be if she responded in the way the steadfaster recommends. For we should expect that if she has a .99 credence in G, as the steadfaster recommends, she will almost certainly have a .99 credence in the truth. After all, her first order evidence – the evidence about the plane itself – strongly supports G. So a steadfaster might claim that hers is the view that would be motivated by an interest in accuracy. After all, the calibrationist recommends responding to the evidence in a way that one should expect, *even a priori*, to be less accurate than the steadfaster's recommendations.

But one might feel like there should be a way to resist this argument. There is a sense, it seems, in which it is precisely *because* Agatha should be concerned about being accurate that she should only adopt a 0.5 credence in G. But what is this sense? How are thinking about the connection between rationality and accuracy if we think that it can be rational to respond to a given body of evidence in a way that one would a priori expect to be less accurate than an alternative response to that evidence?

Perhaps the thought is something like this: the steadfast policy, which has an agent always adopt an attitude that her first order evidence warrants, is not a policy we can expect to always be able to conform to. Indeed, the steadfasting policy tells us to have the attitudes supported by our first order evidence precisely in those cases in which our higher order evidence tells us that we'll likely fail at determining what those attitudes are! Since, we can't expect, in such circumstances, to be able to

[4] For further discussion see White (2009) and Schoenfield (forthcoming).
[5] Some contributors to this project include Joyce (1998), Greaves and Wallace (2006), Moss (2011) and Pettigrew ((2012), (forthcoming)[a] and (forthcoming)[b]).

do what the steadfasting policy recommends, we can't be rationally required to conform to it.

This strikes me as an unpromising route. Consider a body of evidence that includes information to the effect that I am bad at evaluating evidence about the future outcomes of sports matches that I care about a great deal. Say that, for any given credence I adopt on such matters, 50% of the time it turns out that I am a bit overconfident due to wishful thinking, and 50% of the time I am a bit underconfident due to fear of disappointment. I just can't manage to get it right. If my evidence includes this information, then I can't expect to successfully adopt the credences that are supported by my sports related evidence. But that doesn't mean that I'm not rationally required to do so! Beliefs that are unsupported by the evidence due to wishful thinking or fear of disappointment are irrational *even if* the wishful thinkers or fearers can't help themselves and know that they can't. Similarly, I might not be able to expect to rationally evaluate my child's musical abilities. That doesn't mean that there is no attitude that my evidence supports. It just means that I likely won't adopt that attitude. So it can't be that the problem with responding to the evidence in the way the steadfaster recommends is that we can't expect to *succeed* at conforming to her recommendations.

We are left, then, with following puzzling phenomenon: On the one hand, it seems like *especially* when being accurate is extremely important, calibrating makes sense. On the other hand, there's an alternative response to the evidence in these cases – the steadfaster's response – which we should expect to lead to more accuracy than calibrating. I argued that if there's a problem with the steadfaster's view it can't just be that we can't expect to succeed at being steadfast. Being rational can be difficult. It's not a constraint on rational requirements that we should expect to be able to successfully follow them. So in what sense *does* an interest accuracy motivate the calibrationist position?

My goal in this paper is to address this question. I will proceed in two stages. First, I will lay out the accuracy-based framework that is standardly used to motivate rational requirements. I will show that this framework does indeed motivate the steadfaster's position. The argument for this will require generalizing a result proved by Greaves and Wallace for the claim that conditionalizing maximizes expected accuracy. After proving the generalization of their result, I apply it to the case of higher order evidence, and the result that emerges is that, on the standard framework, steadfasting is the rational response to higher order evidence. This has interesting implications for calibrationism and for the accuracy-

driven approach to epistemology. For it turns out that if one wants to pursue the accuracy-based project in the standard way, one takes on the steadfaster's commitments about higher order evidence. On the flip side, the calibrationist can't get on board with the accuracy-driven approach to rational requirements, at least as it is currently conceived.

In the second stage I develop an alternative framework, one on which rational requirements are motivated by an interest in accuracy, but they are not derived from accuracy based considerations in the standard way. Very roughly, what distinguishes the traditional approach from the alternative one is that, on the traditional picture, we're interested in evaluating the expected accuracy of *conforming* to an update-procedure. On the alternative picture that I develop, we instead evaluate the expected accuracy of *planning* to conform to an update procedure. I conclude by listing some advantages of the planning approach and showing how it motivates calibrationism.

## Part I – The Standard Accuracy-Based Approach

The aim of Part I is to provide an accuracy-based argument for steadfasting using the standard accuracy-based approach for deriving rational requirements. To do this, I will appeal to a result proved by Greaves and Wallace: that conditionalization maximizes expected accuracy. As we'll see, however, Greaves and Wallace's result only applies to a limited range of cases. On some ways of conceiving of higher order evidence, cases of higher order evidence won't fall within the scope of cases to which the Greaves and Wallace result applies. So the first step of my argument involves providing a generalization of the result that will apply to cases of higher order evidence. After proving the generalization, I will show how it yields the result that steadfasting maximizes expected accuracy.

As we'll see, the reason that the standard accuracy-based approach delivers the result that we should steadfast, is that the standard approach evaluates a belief revision procedure by calculating the expected accuracy of *conforming* to it. As I mentioned in the introduction, it's not intuitively surprising that we should expect conforming to the steadfaster's recommendations to lead to better results than calibrating. So why bother providing a formal proof?

There are three reasons. First, as I mentioned earlier, the claim that the accuracy-driven approach favors steadfasting has important consequences. It has, for example, the consequence that the calibrationists cannot endorse the accuracy-based approach to deriving rational requirements, at least as it is currently

conceived, and that the accuracy-first proponents are taking on substantive commitments concerning the higher order evidence debate. Since the claim has important consequences, it is worth seeing the argument for it in detail rather than simply relying on one's intuitions. Second, as we'll see, there are some moves the calibrationist can make to resist the thought that steadfasting is more expectedly accurate than calibrating. These moves will become apparent once we lay out the argument in detail. Finally, despite the fact that it's intuitive that steadfasting leads to more accuracy than calibrating, one might wonder *why* this is intuitive: after all, from a calibrationist's perspective, .5 is the credence that's supported by the total body of evidence. (They insist on this!) Why, then, *do* we expect .99 to be more accurate?

In addition to the higher-order evidence based motivations for the argument below, the generalization of Greaves and Wallace that I prove in the course of making this argument is of independent interest, and has its own set of philosophical implications. Since I'll need to use my generalization of the Greaves and Wallace result in arguing that steadfasting maximizes expected accuracy, the first step is to get clear on what exactly Greaves and Wallace show, and then prove the generalization. Doing so requires a bit of setup.

## 2. Some Setup

To begin, I need to introduce some terminology:

### 2.1 Accuracy and Expected Accuracy

**Accuracy** is measured by a scoring rule, **A,** which takes a credence function, *c*, from the set of possible credence functions, C, and a state of the world, s, from a partition of states**,** S, and maps the credence-function/state pair to a number between 0 and 1 that represents how accurate the credence function is in that state.

$$\mathbf{A}: C \times S \rightarrow [0,1]$$

Intuitively, we can think of the accuracy of some credence function as its "closeness to the truth." *c* is maximally accurate if it assigns 1 to all truths and 0 to all falsehoods. It is minimally accurate if it assigns 1 to all falsehoods and 0 to all truths.

If an agent doesn't know what state obtains, she may not be able to calculate the *accuracy* of some credence function, *c*. But, if she is probabilistically coherent, she will be able to calculate the *expected* accuracy of *c*, relative to her probability function *p*.

The **expected accuracy** of credence function *c* relative to probability function *p* is:

$$\mathrm{EA}^{p}(c) = \sum_{s \in S} p(s)\, \mathbf{A}(c, s)$$

That is, the expected accuracy of a credence function, *c*, is the average of the accuracy scores *c* would get in the different states that might obtain, weighted by the probability that those states obtain.

I am going to assume in what follows that the scoring rule that we use to measure accuracy is strictly proper**.** A **strictly proper scoring rule** has the feature that every probability function maximizes expected accuracy relative to itself. In other words, if **A** is strictly proper, then

$$\mathrm{EA}^{p}(c) = \sum_{s \in S} p(s)\, \mathbf{A}(c, s)$$

is maximized when $c = p$.[6]

## 2.2. Learning-experiences and update procedures; Experiments and available acts

Definition 1: Learning Experiences and Update Procedures

A **learning-experience** will be defined as a situation in which an agent is to learn exactly one proposition from a set of propositions $\mathcal{X}$. We will represent a learning-experience by the set of propositions the agent might learn, and we will say that an agent has *undergone* the learning experience, $\mathcal{X}$, when she has learned a member of $\mathcal{X}$.

What is it to learn a proposition? Greaves and Wallace describe learning the proposition P as "gaining the information that P" and they assume that an agent who gains the information that P assigns credence 1 to P. But beyond assigning credence 1 to a proposition, I will make no further assumptions concerning what is involved in learning. (In particular, I will remain neutral about whether one might learn a proposition that is false).

[6] See Greaves and Wallace (2006), Gibbard (2008), Moss (2011), Horowitz (2013) and Pettigrew (forthcoming)[b] for a discussion of the motivation for using strictly proper value functions.

An **update-procedure,** $U$, in response to a learning-experience, $\mathcal{X}$, is a function which assigns to each member of $\mathcal{X}$, a probability distribution, with the intended interpretation that an agent performing $U$ adopts $U(X_i)$ as her credence function if and only if she learns $X_i$.

It will sometimes be convenient to think of $U$ as assigning to each possible *state* (in which the agent learns something) a credence function. So we can define $U(s)$ as $U(X_i)$ where $X_i$ is the proposition in $\mathcal{X}$ that the agent learns in state s. If we let $L(X_i)$ be the proposition that the agent learns $X_i$ upon undergoing the learning experience, we have:

$$U(s) = U(X_i) \text{ where } s \in L(X_i).$$

As we'll see in a moment, an "experiment," in Greaves and Wallace's sense, is just a special kind of learning-experience, and what Greaves and Wallace call an "available-act" is just an update-procedure in response to an experiment. In other words, my notions are generalizations of the notions that Greaves and Wallace use. The difference between the general notions I describe, and the more narrow ones that Greaves and Wallace use will be important in what follows so let me briefly explain what's special about *experiments* and *available acts*.

Definition 2: Experiments and Available Acts

An **experiment,** $\mathcal{E}$ is a learning-experience that satisfies the following two conditions:

PARTITIONALITY: $\mathcal{E}$ is a partition.

FACTIVITY: For all i, if an agent learns $\mathcal{E}_i$, $\mathcal{E}_i$ is true.

The experiment is *performed* when the agent learns a member of $\mathcal{E}$.[7]

Greaves and Wallace's definition of an **available epistemic act *a*** is: "an assignment of a probability distribution to each piece of possible information $\mathcal{E}_j \in \mathcal{E}$, with the intended interpretation that if $a(\mathcal{E}j) = p_j$ then $p_j$ is the probability function that an agent performing act $a$ would adopt as his credence distribution if he received the new information that the actual state was some member of $\mathcal{E}_j$." (611-612). Thus, an available act is just an update-procedure in response to an experiment.

The reason that it's important to talk about learning-experiences, in addition to experiments is as follows: What we're trying to do is figure out the expected

[7] Greaves and Wallace are explicit about their commitment to PARTITIONALITY (p.611), but not to FACTIVITY. It will become clear later why the argument that Greaves and Wallace develop requires FACTIVITY.

accuracy of different ways of revising beliefs in response to new evidence. So far, we have defined the expected accuracy of a *credence function*. But we don't yet have a formal definition of "the expected accuracy of a way-of-revising-beliefs." Since the project is to figure out which way-of-revising-beliefs maximizes expected accuracy we need to figure out exactly what we mean by this. Greaves and Wallace provide an elegant definition for the expected accuracy of an epistemic act in response to an experiment. The problem is that Greaves and Wallace's definition can be applied *only to experiments* (we'll see why this is later). This works perfectly well for their purposes. But we might sometimes be interested in the best way of revising beliefs in response to learning-experiences that *aren't* experiments. As we'll see, some ways of conceiving of higher order evidence will involve considering learning-experiences that aren't experiments. But there are other reasons to be interested in such learning-experiences.

For example, Aaron Bronfman (forthcoming) is interested in cases in which PARTITIONALITY fails. He describes a case in which there is a lottery between three people: A, B and C. Let "A," "B" and "C" name the propositions that the respective people have won the lottery. Suppose A finds out that tomorrow she's going to learn one of the following two propositions:

$$W: \{\sim B, \sim C\}$$

W does not form a partition because it's possible for ~B and ~C to both be true. Thus, A expects to have a learning-experience that isn't an experiment.

What about failures of FACTIVITY? If you think that "learn" is factive, you might think failures of FACTIVITY can never arise, for an agent could never learn a false proposition. But let's aside the semantics of "learn." For various reasons, some philosophers have thought that an agent might have a false proposition as part of her *evidence*.[8] If we're interested in which ways of updating maximize expected accuracy while leaving open the possibility of receiving evidence containing false propositions, we will want to broaden the notion of a learning-experience to allow for such cases.

In sum, if we're interested in how to maximize expected accuracy in cases where the propositions we might learn don't form a partition, or where we leave open the possibility of gaining false information, then we're interested in how to revise beliefs in light of learning-experiences that aren't experiments. Thus, we'll need a notion of "expected-accuracy-of an-update procedure" that applies to

[8] See, for example, Arnold (2013) and Rizzieri (2011).

learning-experiences that aren't experiments. This is the topic of the next subsection.

### 2.3. The Expected Accuracy of Update-Procedures

So what *do* we mean by the expected accuracy of an update-procedure $U$ in response to a learning-experience $\mathcal{X}$? On an intuitive level what we're trying to capture is how accurate we expect an agent who learns a member of $\mathcal{X}$ to be if she conforms to $U$. And recall that, on the intended interpretation, an agent conforms to $U$ if she adopts $U(\mathcal{X}_i)$ whenever she learns $\mathcal{X}_i$.

If we know that some agent is going to undergo a learning-experience $\mathcal{X}$, then we know that she will learn exactly one proposition in $\mathcal{X}$. So consider any state, s, in which our agent learns a proposition in $\mathcal{X}$. (Call this set of states $L(\mathcal{X})$). Let $\mathbf{A}(U(s),s)$ represent the *accuracy score that an agent conforming to U would adopt in s.* (Recall that an agent conforms to $U$ in s if she adopts $U(\mathcal{X}_i)$ in s, where $\mathcal{X}_i$ is the proposition that she learns in s). Now that we have a way of representing how *accurate* an agent conforming to $U$ would be in each state in which she learns a member of $\mathcal{X}$, it is natural to think of the *expected* accuracy of $U$ as the weighted average of the accuracy scores that an agent conforming to $U$ would adopt in each state in which she learns a member of $\mathcal{X}$. This gives us:

Definition 4

The **expected accuracy of an update-procedure *U*** in response to a learning-experience $\mathcal{X}$, relative to a probability function $p$ is:

$$\mathrm{EA}^p(U) = \sum_{s \in L(\mathcal{X})} p(s)\, \mathbf{A}(U(s), s)$$

$$= \sum_{L(\mathcal{X}_i) \in L(\mathcal{X})} \sum_{s \in L(\mathcal{X}_i)} p(s) * \mathbf{A}(U(\mathcal{X}_i)), s)$$

This quantity represents a weighted average of accuracy scores that that would result from an agent adopting $U(\mathcal{X}_i)$ as her credence function whenever she learns $\mathcal{X}_i$.

### 2.4. A Lemma

I will now prove the following lemma:

Lemma:
If $\mathcal{E}$ is an experiment and $A$ is an update-procedure in response to $\mathcal{E}$:

$$\mathrm{EA}^{p}(A) = \sum_{\mathrm{L}(\mathcal{E}_i)\in \mathrm{L}(\mathcal{E})} \sum_{s\in \mathrm{L}(\mathcal{E}_i)} p(s)^{*}\,\mathbf{A}(A(\mathcal{E}_i)), s) = \sum_{\mathcal{E}_i \in \mathcal{E}} \sum_{s\in \mathcal{E}_i} p(s)^{*}\,\mathbf{A}(A(\mathcal{E}_i)), s)$$

Note that the first (leftmost) double sum is just the definition of the expected accuracy of an update-procedure. The second double sum is just like the first one except that rather than summing over the $\mathrm{L}(\mathcal{E}_i)$, we're summing over the $\mathcal{E}_i$.

Proof
The proof will proceed by showing that if $\mathcal{E}$ is an experiment, then for all propostitions $\mathcal{E}_i \in \mathcal{E}$:

$$\mathcal{E}_i \leftrightarrow \mathrm{L}(\mathcal{E}_i)$$

Given this biconditional, there is no harm in replacing the "$\mathrm{L}(\mathcal{E}_i)$" that features in the *definition* of the expected accuracy of an update-procedure with "$\mathcal{E}_i$."

Here is the argument for the biconditional:

Let $\mathcal{E}$ be any experiment. This means that $\mathcal{E}$ is a learning-experience that satisfies FACTIVITY and PARTITIONALITY.

FACTIVITY entails the right-to-left direction of the biconditional: $\mathrm{L}(\mathcal{E}_i) \rightarrow \mathcal{E}_i$. For FACTIVITY says that if an agent learns $\mathcal{E}_i$, $\mathcal{E}_i$ must be true.

What about the left-to-right direction? If PARTIONALITY holds then exactly one proposition in $\mathcal{E}$ is true. Since the agent will learn one proposition in $\mathcal{E}$, and (due to FACTIVITY) it will be a true proposition, she will have to learn *the* one true proposition in $\mathcal{E}$. So if $\mathcal{E}$ forms a partition, we know that the $\mathcal{E}_i$ that is true is the proposition that she will learn. This gives us: $\mathcal{E}_i \rightarrow \mathrm{L}(\mathcal{E}_i)$.

Since Greaves and Wallace assume FACTIVITY and PARTITIONALITY they can simply *define* the expected accuracy of an act in response to an experiment as the average accuracy scores that would result from adopting $A(\mathcal{E}_i)$ whenever $\mathcal{E}_i$ *is true*. And this, indeed, is what they do. Their definition of expected accuracy corresponds to the double sum on the right-hand side of the lemma. But it's important to realize that they *wouldn't* define expected accuracy this way if they weren't assuming FACTIVITY and PARTITIONALITY. This is because, without these assumptions, the double sum on the right does not represent a weighted average of the scores that would result from an agent performing act $A$. For recall that Greaves and Wallace, in defining an available act, say that the intended interpretation is that an agent performs act $A$ in response to $\mathcal{E}$ if she adopts $A(\mathcal{E}_i)$ as her credence function if and only if *she learns* $\mathcal{E}_i$ (p.612). But if $\mathcal{E}_i$ might true, even though the agent doesn't learn it, or if she learns it but it's not true, then an agent performing $A$ would *not* adopt $A(\mathcal{E}_i)$ if and only if $\mathcal{E}_i$ is true. Thus, it is only if FACTIVITY and PARTITIONALITY are assumed that the double sum on the right accurately represents the expected accuracy of the credences that result from an agent performing $A$.

**(f) Summing Up**

The purpose of this section was to develop a precise definition of the notion of the expected accuracy of an update-procedure in response to a learning-experience. Although Greaves and Wallace provide a definition for the expected accuracy of an *act* in response to an *experiment*, this definition won't apply to learning-experiences that aren't experiments. Examples of such cases are cases in which we think we might learn one proposition from a set of propositions that doesn't form a partition (as in Bronfman's lottery case), or cases in which we think we might get false information. Since such cases can't be regarded as experiments, the Greaves and Wallace framework does not apply to them.

I defined the expected accuracy of an update-procedure as the weighted average of the accuracy scores that would result form an agent conforming to the update-procedure. Since, on the intended interpretation, an agent conforming to $U$ adopts $U(\mathcal{X}_i)$ whenever she learns $\mathcal{X}_i$, to calculate this quantity we need to average the accuracy scores that would result from an agent adopting $U(\mathcal{X}_i)$ whenever $\mathcal{X}_i$ is the proposition learned. I then showed that if an update-procedure is an experiment, this will be equivalent to averaging the accuracy scores that would result from an agent adopting $U(\mathcal{X}_i)$ whenever $\mathcal{X}_i$ is *true*. This gives us Greaves and Wallace's definition of the expected accuracy of an act. Thus, my framework, in

terms of update-procedures and learning-experiences, is just a generalization of the framework developed by Greaves and Wallace.

We can now make the question of which update-procedure maximizes expected accuracy in HYPOXIA more precise. Suppose that Aisha, on Sunday, is considering how she should respond to the different bodies of evidence she might get when she's on her plane on Monday. To do this, Aisha considers the set of propositions she might learn on her flight. I will discuss exactly what set this is in the following sections, but for now, let's just call her learning-experience, whatever it is, "FLIGHT." We can now consider different update-procedures one might have in response to FLIGHT and ask which of these has highest expected accuracy. That is, we consider different functions $U$, from the propositions, $\mathcal{F}_i$, in FLIGHT to credence functions. We then ask which function $U$ maximizes the expected accuracy of the credences that would result from Aisha adopting $U(\mathcal{F}_i)$ whenever she learns $\mathcal{F}_i$. This is the question that will be the focus of Part II of the paper.

However, to answer this question we will first have to generalize the result that Greaves and Wallace prove: the claim that *conditionalizing* on the proposition one learns on the basis of an experiment maximizes expected accuracy. In the next section, I prove a more general result: The update-procedure that maximizes expected accuracy in response to *any* learning-experience is one in which an agent who learns $\mathcal{X}_i$ conditionalizes on the proposition *that she learned* $\mathcal{X}_i$ *upon undergoing the learning-experience.* In cases in which the learning-experience is an experiment, this amounts to the same thing as conditionalizing on $\mathcal{X}_i$.

## 3. The Greaves and Wallace Result and its Generalization

Greaves and Wallace (2006) present an argument for the claim that conditionalization on the proposition that one learns is the update-procedure that maximizes expected accuracy in response to an *experiment* if one's scoring rule is strictly proper. One conditionalizes on a proposition, Q, if

$$p_{new}(\cdot) = p_{old}(\cdot \mid Q)$$

Where

$$p(A|B) = p(A\&B) \,/\, p(B)$$

We can think of the argument for this claim as involving two steps. First, there is a purely formal result that demonstrates that plugging in certain values in certain quantities maximizes other quantities. Second, there is an argument *from* this formal result to the claim that, given our understanding of update-procedures,

expected accuracy of update-procedures, learning, and experiments, the update-procedure (or available act) that maximizes expected accuracy in response to an experiment is the one that has the agent conditionalize on the proposition she learns. It will be important to keep these two steps separate. I will call the purely formal result that can be extracted from Greaves and Wallace's paper "G&W":

> **G&W**: Take any partition of states $\mathcal{P}$: $\{\mathcal{P}_1...\mathcal{P}_n\}$ and consider the set of functions, $\mathcal{G}$, that assign members of $\mathcal{P}$ to probability functions. The member of $\mathcal{G}$, $F$, that maximizes this quantity:
>
> $$\sum_{\mathcal{P}_i \in \mathcal{P}} \sum_{s \in \mathcal{P}_i} p(s)^* \mathbf{A}(F(\mathcal{P}_i), s)$$
>
> is:
>
> $$F(\mathcal{P}_i) = \text{Cond} = p(\cdot \mid \mathcal{P}_i)$$
>
> (where **A** is strictly proper).

**G&W** can be used to derive Greaves and Wallace's claim about experiments:

> **CondMax**: Suppose you know that you are going to perform an experiment, $\mathcal{E}$.[9] The update-procedure that maximizes expected accuracy in response to $\mathcal{E}$, relative to probability function $p$, is the update-procedure that assigns, to each $\mathcal{E}_i$, $p(\cdot \mid \mathcal{E}_i)$.

The argument from **G&W** to **CondMax,** using our generalized framework is simple.

> Proof of CondMax:
>
> (1) The expected accuracy of an update-procedure $U$ in response to an experiment E, relative to a probability function $p$ is:

[9] For the purposes of this paper I am assuming that if an agent knows P then she is rationally certain that P and P is true. If you don't like this assumption simply substitute "rationally certain that P and P is true" for "knows P." Greaves and Wallace are considering cases where one assigns credence 1 to the proposition *that one will learn exactly one (true) proposition from a relevant partition*. I will provide a more generalized result below. But you might wonder: "what if I'm not certain that I will learn *anything*? Perhaps my laboratory will blow up and the experiment won't even be performed!" We could derive even further generalizations from **G&W** to allow for such possibilities. However, since doing so won't deliver any additional benefits when it comes to discussing higher order evidence, which is my primary focus, I will not provide further generalizations here.

(*) $$\sum_{\mathcal{E}_i \in \mathcal{E}} \sum_{s \in \mathcal{E}_i} p(s)^* \mathbf{A}(U(\mathcal{E}_i), s)$$

(from Lemma).

(2) The value of $U$ that maximizes (*) is $U$=Cond ($\mathcal{E}_i$).

(This follows from **G&W** and the fact that $\mathcal{E}$ is a partition)

(3) The update-procedure $U$ that maximizes expected accuracy in response to an experiment $\mathcal{E}$ is $U$=Cond($\mathcal{E}_i$). That is, the update-procedure that maximizes expected accuracy is the one that has the agent conditionalize on the member of $\mathcal{E}$ that she learns.
(This follows from (1) and (2)).

But what about cases in which our learning-experiences won't be experiments? What update-procedure maximizes expected accuracy in those cases? Here is the answer:

**Generalized CondMax:** Suppose you know that you are going to undergo a learning-experience, $\mathcal{X}$. The update-procedure that maximizes expected accuracy in response to $\mathcal{X}$, relative to probability function $p$, is the update-procedure that assigns, to each $\mathcal{X}_i$, $p(\cdot|L(\mathcal{X}_i))$ where $L(\mathcal{X}_i)$ is the proposition that the agent learns $\mathcal{X}_i$ upon undergoing the learning-experience.

Proof of Generalized CondMax:
Recall that the expected accuracy of an update-procedure, $U$, in response to a learning-experience, $\mathcal{X}$ is defined as:

(#) $$\sum_{L(\mathcal{X}_i) \in L(\mathcal{X})} \sum_{s \in L(\mathcal{X}_i)} p(s)^* \mathbf{A}(U(\mathcal{X}_i)), s)$$

We are aiming to show is that # is maximized when $U(\mathcal{X}_i)$ = Cond($L(\mathcal{X}_i)$). So, suppose for reductio that this is false, that is, that there exists a function, $U^*$, such that:

$$\sum_{L(\mathcal{X}_i)\in L(\mathcal{X})} \sum_{s\in L(\mathcal{X}_i)} p(s)^* \mathbf{A}(U^*(\mathcal{X}_i)), s) \quad > \quad \sum_{L(\mathcal{X}_i)\in L(\mathcal{X})} \sum_{s\in L(\mathcal{X}_i)} p(s)^* \mathbf{A}(\text{Cond}(L(\mathcal{X}_i)), s)$$

Now, define μ(L(Xi)) as $U^*(\mathcal{X}_i)$.[10] It follows that:

$$\sum_{L(\mathcal{X}_i)\in L(\mathcal{X})} \sum_{s\in L(\mathcal{X}_i)} p(s)^* \mathbf{A}(\mu(L(\mathcal{X}_i)), s) \quad > \quad \sum_{L(\mathcal{X}_i)\in L(\mathcal{X})} \sum_{s\in L(\mathcal{X}_i)} p(s)^* \mathbf{A}(\text{Cond}(L(\mathcal{X}_i)), s)$$

But this is impossible, because it follows from **G&W** that the quantity:

(##) $$\sum_{L(\mathcal{X}_i)\in L(\mathcal{X})} \sum_{s\in L(\mathcal{X}_i)} p(s)^* \mathbf{A}(F(L((\mathcal{X}_i))), s)$$

is maximized when $F$ = Cond(L(Xi)). Thus, there cannot exist a μ that satisfies the inequality above. Contradiction.

Here is the lesson to be learned from CondMax and its generalization: the update-procedure that maximizes expected accuracy in response to *any* learning-experience is the one in which an agent who learns $\mathcal{X}_i$ conditionalizes on the proposition *that she learned* $\mathcal{X}_i$ upon undergoing the learning experience. The reason that, in the case of experiments, conditionalizing on the proposition that one learns maximizes expected accuracy is that, in these special cases, the agent knows that she will learn a proposition $\mathcal{X}_i$ if and only if $\mathcal{X}_i$ is true. Thus conditionalizing on $\mathcal{X}_i$ amounts to the very same thing as conditionalizing on $L(\mathcal{X}_i)$.[11]

The fact that, when $\mathcal{X}$ and $L(\mathcal{X})$ come apart, the update procedure that maximizes expected accuracy has us conditionalize on a proposition in $L(\mathcal{X})$ may be both surprising and disconcerting. Here's why: Conditionalizing on $L(\mathcal{X}_i)$ involves assigning credence 1 to $L(\mathcal{X}_i)$. This means that if you think that the update-procedure that maximizes expected accuracy is the *rational* update procedure, it

[10] How do we know that there is such a $\mu$? Since there is a bijection between the $\mathcal{X}_i$ and the $L(\mathcal{X}_i)$ there exists an inverse of $L(\mathcal{X}_i)$, which we'll call "$L^-(\mathcal{X}_i)$," such that $L^-(L(\mathcal{X}_i)) = \mathcal{X}_i$. We can then let $\mu$ $(L(\mathcal{X}_i))$ be $U^*$ composed with $L^-$. Thus: $\mu$ $(L(\mathcal{X}_i)) = U^*(L^-(L(\mathcal{X}_i)) = U^*(\mathcal{X}_i)$.

[11] Indeed, the mistake of thinking that conditionalizing *on the proposition one learns* maximizes expected accuracy, when the possible propositions one might learn don't form a partition is, as Aaron Bronfman (forthcoming) points out, the fallacy involved in the famous Monty Hall problem. (The right answer to Monty Hall results from conditionalizing *not* on the proposition that there's a goat behind door 2, but rather on the proposition that *one learned that there is a goat behind door 2*).

follows that any agent who learns $\mathcal{X}_i$ is rationally required to be *certain* that she learned $\mathcal{X}_i$. But if you reject KK – the principle that whenever an agent knows P, she is in a position to know that she knows P – the claim that an agent who learns P is rationally required to be certain that she learned P may be tough to swallow.

There are different ways one might respond to this concern.

One option is to conclude that agents *are* rationally required to be certain that they learned what they learned. Another is to conclude that that the accuracy-driven approach should be abandoned since KK is false. But I suspect that many will find such responses unattractive, and, since there are less radical alternatives available, they are also unnecessary.

The second option is to follow Bronfman's suggestion for how to deal with failures of partitionality. He suggests that in cases in which it is not plausible that an agent should be certain about which proposition she learned, we should simply remove from the set of potential update-procedures on $\mathcal{X}$ those which require assigning credence 1 to $L(\mathcal{X}_i)$ upon learning $\mathcal{X}_i$. We then sift through the ones that remain and figure out which of *those* maximizes expected accuracy.

The third strategy is one that I will pursue at the end of this paper. The idea is that we *can* derive rational requirements from accuracy based considerations, but not in the manner recommended by the standard framework.

For now, though, I will continue to follow the standard picture. And to sidestep questions about whether agents, in general, are always in a position to be rationally certain that they learned what they learned, I will just stipulate that Aisha, in the hypoxia case, is rationally certain that she learns what she learns. The question of what Aisha should do if she learns that she's impaired but she doesn't become certain that she learned this is a question for another day.

Here is where we are: I have defined the notion of a learning-experience and an update procedure in response to a learning-experience broadly enough to allow for cases in which (a) the propositions that we might learn don't form a partition, and (b) we might learn a false proposition. I then showed that the formal result that Greaves and Wallace prove can be used to derive two theorems. Generalized CondMax tells us that the accuracy optimizing update-procedure in response to any learning-experience, $\mathcal{X}$, has us conditionalize on the proposition $L(\mathcal{X}_i)$ upon learning $\mathcal{X}_i$. CondMax tells us that in the special case in which we are certain that our learning-experience will be an experiment this will be equivalent to conditionalizing on $\mathcal{X}_i$ upon learning $\mathcal{X}_i$.

We are now in a position to apply the generalized framework to the case of higher order evidence.

## 4. What Aisha Learns in HYPOXIA

It is important to realize that, in the framework we're working with, we consider the expected accuracy of various update procedures *before* undergoing the learning-experience. (I will discuss the reasons for this later). Since Aisha flies her plane on Monday, let's imagine that, on Sunday, she is considering which update-procedure maximizes expected accuracy in response to the learning-experience that will take place on Monday.

To figure out which update-procedure maximizes expected accuracy in Aisha's case we must first determine which set represents the learning-experience that she will undergo on the flight. Then we can determine, using CondMax, or Generalized CondMax, which update-procedure maximizes expected accuracy in response to that learning-experience.

So let's begin with Step 1: determining which set (on Sunday) represents Aisha's learning-experience on Monday.

### 4.1. Being Impaired *on Monday*

You might think that what Aisha learns while flying the plane is the de dicto proposition: *E and Aisha's reasoning is impaired on Monday.* Let us suppose that, on Sunday, Aisha knows that on Monday she will learn both whether E and whether she is impaired on Monday, and that what she will learn will be true. That is, Aisha on Sunday, knows that she'll be performing an *experiment* on Monday in which she'll learn one of the following four propositions:

(1) E and Aisha's reasoning is impaired on Monday.

(2) E and Aisha's reasoning is not impaired on Monday.

(3) ~E and Aisha's reasoning is impaired on Monday.

(4)~E and Aisha's reasoning is not impaired on Monday.

Let $H_m$ be the proposition that Aisha's reasoning is impaired on Monday. We can then represent the four propositions Aisha might learn in "flight experiment" as:

$\mathcal{F}$: {$EH_m$, $E{\sim}H_m$, $\sim EH_m$, $\sim E{\sim}H_m$}

Since, on this picture, Aisha's learning-experience on Monday will be an experiment (it satisfies FACTIVITY and PARTITIONALITY), CondMax tells us that, on

Sunday, Aisha should regard conditionalizing on the members of $\mathcal{F}$ as the update procedure that maximizes expected accuracy.

But conditionalizing on the propositions in $\mathcal{F}$ yields *steadfastness*. The reason for this is that, on Sunday, Aisha will regard the quality of her reasoning capacities on *Monday* to be completely irrelevant to the question of whether there will be enough gas in the tank on the supposition that various facts about the plane obtain. So, on Sunday, Aisha's credence that G conditional on E will be the same as Aisha's credence that G conditional on E *and Aisha's reasoning is impaired on Monday*. This means that, if $p_s$ represents Aisha's probabilities on Sunday, we'll have

- $p_s(G|E) = 0.99$
- $p_s(G|EH_m) = 0.99$

But then, on the assumption that what Aisha learns in HYPOXIA is $EH_m$, the accuracy optimizing act is one which assigns *0.99* to G in HYPOXIA, since that is the credence in G that would result from conditionalizing.

David Christensen (a calibrationist) makes a similar point (though not in terms of conditional probabilities) in his (2010). He writes:

> "So it seems that the [higher order evidence] about my being drugged produces a mismatch between my current confidence that H is true on the supposition that I will learn certain facts, and the confidence in H that I should adopt if I actually learn those facts" (200).

### 4.2. Being Impaired *Now*

The purpose of this subsection is to explore a suggestion made by Christensen about an alternative way of thinking about Aisha's evidence. On this suggestion, rather than thinking of Aisha's evidence on Monday as *E and Aisha is impaired on Monday*, we think of her evidence on Monday as the essentially indexical proposition: *E and I'm impaired now.* The thought is that just as the practical import of "S is trailing sugar from her cart at time t" is different from the practical import of "*I* am trailing sugar from my cart *now*" the *epistemic* import of "S's reasoning is impaired at time t" is different from the epistemic import of "my reasoning is impaired now."

Let "$H_{now}$" be the proposition that my reasoning is impaired *now*. We can represent $H_{now}$ as the set of centered worlds in which the center's reasoning is impaired. The suggestion hinted at by Christensen is that, while perhaps conditionalizing on $H_m$ doesn't yield calibrationist results, conditionalizing on $H_{now}$

does. This is because, the calibrationist might claim, even on *Sunday*, Aisha's credence in G, on the supposition that "E and my reasoning is impaired *now*" should be 0.5. That is, the calibrationist may claim that Aisha's conditional probabilities on Sunday should look like this:

- $p_s$ (G|E) = 0.99
- $p_s$ (G|$EH_{now}$) = 0.5.

These conditional probabilities at least *appear* to be compatible with thinking of calibrating as a kind of conditionalizing.

Recall that the reason that we're interested in whether we can model the calibrationist as a conditionalizer is that we already know that conditionalizing in response to an experiment maximizes expected accuracy. Now, suppose that the calibrationist can make the case that the four propositions Aisha might learn on Monday are:

$$\mathcal{F}^*: \{EH_{now}, E{\sim}H_{now}, {\sim}EH_{now}, {\sim}E{\sim}H_{now}\}$$

If Aisha's learning experience on Monday should be represented by $\mathcal{F}^*$, then, if she knows that her learning-experience will be an experiment, CondMax tells us that she should, on Sunday, regard conditionalizing on the members of $\mathcal{F}^*$as the update-procedure that maximizes expected accuracy. This might give the calibrationist everything she needs to defend her approach to higher order evidence. It *is*, she might claim, the approach that maximizes expected accuracy. One simply has to be clear about the fact that the evidence in higher order evidence cases is essentially indexical.

There are a variety of worries one might have with this strategy. For example, one may be skeptical that conditionalization is the right way to respond to indexical evidence more generally. But I'm going to set these worries about the best theory of *de se* updating aside. For now, I'm just interested in focusing on a very narrow question: which update-procedure should Aisha, on Sunday, regard as the one that maximizes her expected accuracy on Monday concerning the proposition G. As we'll see, even if, in this particular case, the calibrationist move *can* be modeled as a kind of conditionalization, it won't be the kind that maximizes expected accuracy. This is because if, on Sunday, Aisha thinks that her learning-experience on Monday should be represented by $\mathcal{F}^*$, she shouldn't think that her learning experience-will be an experiment. Thus, conditionalizing on the members of $\mathcal{F}^*$ won't maximize expected accuracy.

Why is this? Recall that for a learning-experience, $\mathcal{E}$, to be an experiment, two conditions must be satisfied:

FACTIVITY: For all i, if an agent learns $\mathcal{E}_i$, $\mathcal{E}_i$ is true.

PARTITIONALITY: $\mathcal{E}$ is a partition.

On the face of it, it may appear that Aisha *should* think that her learning-experience on Monday will satisfy both of these conditions. After all, the propositions in $\mathcal{F}^*$ do form a partition, and we're assuming that she'll only learn the truth. So what's the problem?

The problem is that even though Aisha knows that, on Monday, what she will learn will be true, if the Monday learning-experience is represented by $\mathcal{F}^*$, on *Sunday*, FACTIVITY will be false. For FACTIVITY to hold on Sunday, Aisha would have to think that if she learns $H_{now}$ upon undergoing her learning-experience (which takes place on Monday), then $H_{now}$ is true. But on Sunday, she doesn't think that if she will learn that she is impaired on Monday, then she is impaired on Sunday! So it's simply not true, on Sunday, that for every $\mathcal{F}_i$ in $\mathcal{F}$, Aisha is certain that if she will learn $\mathcal{F}_i$ upon undergoing the learning-experience, $\mathcal{F}_i$ is true. And this is all that is required for FACTIVITY to fail. Since FACTIVITY fails, CondMax *doesn't* yield the result that conditionalizing on the members of $\mathcal{F}^*$ is the update-procedure that maximizes expected accuracy even if her Monday learning-experience *is* accurately represented by $\mathcal{F}^*$.

To see this point more clearly it might be helpful to recall the important role that FACTIVITY played in the argument for the claim that conditionalizing maximizes expected accuracy. Recall that, on the intended interpretation of an update-procedure (or available act), an agent adopts $U(\mathcal{X}_i)$ whenever she learns $\mathcal{X}_i$. Thus, the expected accuracy of an agent performing $U$ is a weighted average of the accuracy scores that would result from adopting $U(\mathcal{X}_i)$ whenever the agent learns $\mathcal{X}_i$:

$$EA^p(U) = \sum_{s \in L(\mathcal{X})} p(s)\, \mathbf{A}(U(s), s)$$

$$= \sum_{L(\mathcal{X}_i) \in L(\mathcal{X})} \sum_{s \in L(\mathcal{X}_i)} p(s)^*\, \mathbf{A}(U(\mathcal{X}_i)), s)$$

It is only acceptable to substitute the $L(\mathcal{X}_i)$ with $\mathcal{X}_i$, in the definition above, if one thinks that one will learn $\mathcal{X}_i$ (and hence adopt $U(\mathcal{X}_i)$) if and only if $\mathcal{X}_i$ is true. And conditionalizing *on the learned proposition* only maximizes expected accuracy if this substitution is acceptable. But now note that in the case of essentially indexical evidence like $H_{now}$ this substitution isn't acceptable. Aisha, on Sunday, won't think

that she will learn $H_{now}$ if and only if she is impaired *now*. Since the credence Aisha will form on Monday, if she performs *U*, will depend on what she learns on *Monday*, it would be a huge mistake to calculate expected accuracy in a way that presupposed that her credences on Monday will depend on her degree of impairment *on Sunday*. This is why, on Sunday, $L(H_{now})$ can't be substituted for $H_{now}$, even though Aisha knows that the proposition that she will learn on Monday will be true (on Monday). Since this substitution can't be performed, CondMax doesn't apply.

So what is the update-procedure that maximizes expected accuracy if Aisha's learning experience on Monday should be represented by $\mathcal{F}$*? Generalized CondMax tells us that the accuracy optimizing update-procedure will be one in which Aisha conditionalizes on the propositions in $L(\mathcal{F}^*)$:

$L(\mathcal{F}^*)$: {Aisha learns $EH_{now}$ upon performing the experiment
Aisha learns $E{\sim}H_{now}$ upon performing the experiment
Aisha learns $\sim EH_{now}$ upon performing the experiment
Aisha learns $\sim E{\sim}H_{now}$ upon performing the experiment}

This means that the accuracy optimizing update procedure is one in which, if she learns $EH_{now}$ she adopts the credence function that results from conditionalizing on the proposition *that she learned* $EH_{now}$. But now note that, given what Aisha knows about her learning-experience on Monday, $L(\mathcal{F}^*)$ is equivalent to $\mathcal{F}$. Aisha will learn $EH_{now}$ upon performing the experiment if and only if $EH_m$, she will learn $E{\sim}H_{now}$ if and only if $E{\sim}H_m$, and so forth. Since we've already established that conditionalizing on the members of $\mathcal{F}$ yields steadfastness, it follows that conditionalizing on the members of $L(\mathcal{F}^*)$ also yields steadfastness.

Thus, even if the calibrationist thinks that Aisha receives centered evidence on Monday, the update-procedure which maximizes expected accuracy *in response to that centered evidence* is not the one which has her conditionalize on the centered propositions. It is the one that has her conditionalize on the proposition that she learned a centered proposition. And this, as we saw, yields steadfastness.

At this point, the calibrationist may respond as follows: "All you've shown me is that, on *Sunday*, I should think that the best update-procedure *for Monday* is the steadfast one. But you haven't shown me that, *on Monday*, I should think that the best update-procedure is the steadfast one. For, on Monday, $\mathcal{F}$* *does* accurately represent an experiment that takes place on Monday. So I can think *on Monday*, upon learning $EH_{now}$, that having the probabilities that result from conditionalizing on $EH_{now}$ *does* maximize expected accuracy. And don't you think that, on Monday, I

should be more concerned with what my Monday credences tell me maximizes expected accuracy, than my Sunday ones?"

There's a sense in which the calibrationist is exactly right. Call Aisha's probablility function on Monday "$p_m$." And say that her total evidence on Monday is M. Assuming she assigns 1 to M, the credence function that Aisha should regard as maximizing expected accuracy on Monday is indeed, $p_m(\cdot|M)$. But that's just because $p_m(\cdot|M)$ equals $p_m$, and since we're using strictly proper scoring rules, every probability function maximizes expected accuracy relative to itself. Thus, since $p_m$ maximizes expected accuracy relative to $p_m$, $p_m(\cdot|M)$ also maximizes expected accuracy relative to $p_m$.

What this demonstrates is that accuracy based considerations don't deliver any particularly interesting results about how to respond to higher order evidence (or really, any sort of evidence) if we think of the question as follows: what credence function maximizes expected accuracy *relative to the credence function that you have adopted* upon receiving the higher order evidence? No matter what update-procedure you use, *once you've updated*, your new probability function will regard itself as maximizing expected accuracy.

The project of thinking about which update-procedures maximize expected accuracy is, essentially, a diachronic one. At very least, it is one that spans more than one probability function. The question is: relative to $p_1$ which probability function should I hope to adopt if I learn E? Suppose the answer is $p_2$. Once you've learned E, and you've adopted some other probability function, say, $p_3$, it is always open to you to say: "$p_3$ maximizes expected accuracy relative to itself! It's true that, from the perspective of $p_1$, $p_2$, rather than $p_3$ is the function I would have hoped to adopt if I learned E. But why should I care how things look from the perspective of $p_1$ if my current perspective is that of $p_3$?" I'm not going to try to answer this challenge. All I will point out is that the thought that underlies the standard accuracy-based approaches for deriving rational requirements, like the argument for the claim that it's rational to conditionalize, is that it's rational to have the credences that some prior probability function would have regarded as maximizing expected accuracy. If this is the picture of how rational requirements and expected accuracy considerations relate to one another, steadfastness wins. In the next section of the paper I will argue that there's an alternative way of deriving rational requirements from accuracy-based considerations which does, indeed, favor calibrationism.

Part II – An Alternative Approach

## 5. An Alternative Approach

Let's set aside rationality for a moment and think about the following *activity*: *deliberating about what to believe*. Deliberating about what to believe is something we do all the time. We may think about what to believe in our current situation, but we also sometimes consider what to believe in possible or future situations. A scientist, for example, may wonder what to conclude about theory T if her experiment delivers result R. I might consider how confident to be in my diagnosis of what's wrong with my car if I learn that my friend (an amateur mechanic, herself) disagrees with me. Sometimes, the outcome of this deliberation is that we settle for ourselves the question of what to believe in the circumstance in question. Perhaps, for example, I settle on suspending judgment if learn that my friend disagrees with me about the car. When we settle our deliberations about what to believe in a certain way, I will say that we have made a *doxastic plan*.[12] We can think of belief-revision procedures as representing doxastic plans.

On the standard accuracy-based picture, we evaluate an update procedure by evaluating the expected accuracy of the credences that result from *conforming* to that procedure. But if we're deliberating about what doxastic plan to adopt, we might, instead, be interested in how accurate we expect to be as a result of *planning* to update in a certain way. Why is this? Let me begin with an illustration from practical planning.

Suppose that I am planning my vacation and I am considering two possibilities: spending my vacation in Paris or spending my vacation on the moon. Clearly, vacationing on the moon would be more exciting than vacationing in Paris. Nonetheless, the moon plan is worse than the Paris plan. Why is this?

One might claim that I simply can't plan to go to the moon because I don't believe that I will, or can, conform to the moon plan. But such an explanation would appeal to controversial principles about planning, such as the principle that says that in order to plan to $\phi$ you must believe that you can, or will, $\phi$. These are principles which I would like to remain neutral about. Therefore, I prefer to appeal to the following very minimal thing that we can say about the moon plan, which suffices to explain its badness: even supposing that I *can* make the plan, I can't

[12] The claim that we engage in doxastic planning doesn't presuppose voluntarism about belief. See Schafer (2014) and Gibbard (20013) for further discussion on doxastic planning.

expect anything good to come of it. (In fact, I can probably expect something bad to come of it, like spending my vacation moping around at home, feeling defeated by my failure). On the other hand, if I plan to go to Paris, the likely result is that I go to Paris and have a terrific time. So while I can expect that the result of *conforming* to the moon plan will be better than the result of conforming to the Paris plan, I can also expect that the result of *making* the moon plan (again, assuming I *can* make such a plan) will be worse than the result of making the Paris plan.

We can apply the distinction between evaluating the results of *conforming* to a plan and evaluating the results of *making* a plan to doxastic planning. The crucial difference between these two activities is that, in the latter case, we can take into account the possibility that we'll fail, and how bad such failures will be. With this in mind, I will argue that *planning* to calibrate does as well, or better, expected accuracy wise, than *planning* to update in any of the ways that have been proposed in the literature.[13]

First, let's define the expected accuracy of *planning* to *update in in accord with U*. Let T be a partition over the possibilities in which the agent plans to update in accord with *U*. And let T be sufficiently fine grained so as to determine for each t $\in$ T: (a) what credence function the agent adopts in t and (b) how accurate that credence function is. Let $PL^U(t)$ = the credence function that the agent adopts in t. We can define the expected accuracy of planning to *U*, relative to a probability distribution *p* over T, as:

$$EA^p(PL^U) = \sum_{t \in T} p(t)\, \mathbf{A}(PL^U(t), t)$$

We can now ask: what's the expected accuracy of planning to steadfast?

If you plan to steadfast, then you plan to adopt the credences supported by your first order evidence, even if you have higher order evidence suggesting that you're impaired. In the hypoxia case, this means that you plan to assign a .99 credence to the proposition that the first order evidence best supports. (And, let's suppose, a .01 credence to its negation). Once again, you might think that one simply *can't* make such a plan because one can't rationally believe that one will

[13] Steel (ms.) is (independenly) developing a similar approach to defend a conciliatory response to peer disagreement.

conform to it. But, as I said earlier, I do not want to rely on any controversial principles about planning, nor do I need to.

Instead, let's assume for the sake of argument that one *can* plan to steadfast and focus on the expected results of making such a plan. The important thing to realize is that, even if you *plan* to steadfast in cases in which you're hypoxic, you should expect that the judgments[14] you will actually make on the basis of the first order evidence will be correct only 50% of the time. Thus, you should expect that if you plan to steadfast in cases in which you're hypoxic, 50% of the time you'll assign a 0.99 credence in the truth (and .01 to the falsehood) and get quite a high accuracy score and 50% of the time you'll assign a .99 credence to the falsehood (and .01 to the truth) and get quite a low accuracy score.[15]

To be more precise about this, let's restrict our attention to your credences in G and ~G. And, for any r $\in$ [0,1], let *r* be the probability function that assigns credence r to G and credence 1-r to ~G. Thus, *.99* is the function that assigns a .99 credence to G and a .01 credence to ~G.

Assuming that our accuracy measure assigns the same value to having credence r in the truth, whether the truth is G or ~G, we can represent the accuracy score that one would get for having a credence r in the truth about whether G (and 1-r in the falsehood) as $\mathbf{A}(r, G)$. Similarly, assuming that the accuracy measure assigns the same value to having a credence r in the falsehood, whether the truth is G or ~G, we can represent the accuracy score that one would get for having credence r in the falsehood (and a 1-r in the truth) as $\mathbf{A}(r, \sim G)$. In sum:

| | |
|---|---|
| Score for r in the truth and 1-r in the falsehood | $\mathbf{A}(r, G)$ |
| Score for r in the falsehood and 1-r in the truth | $\mathbf{A}(r, \sim G)$ |

[14] The proposition that you judge is the proposition from the relevant partition (in our case: {G,~G}) that you were or would be most confident in on the basis of the first order evidence alone. I borrow this term from Horowitz and Sliwa (ms.) and Weatherson (ms.).

[15] I am assuming here that the result of your planning to steadfast will involve assigning a .99 credence to the proposition that you judge (see previous note) on the basis of the first order evidence. You might question whether this is actually what an agent who planned to steadfast would do. Perhaps, rather than using her judgments, she would think: "Oh my! My plan was to assign a .99 credence to the proposition that the first order evidence supports, but, because I'm hypoxic, I don't know what proposition this is! I better just stick to my initial 0.5 credence." If this is what one expects the result of planning to steadfast will be, then the expected result of planning to steadfast will be the same as the expected result of planning to calibrate. So, as a plan, steadfasting would have no advantage over calibrating.

Since, as noted above, if you plan to steadfast you should expect that 50% of the time you'll assign .99 to the truth and 50% of the time you'll assign .99 to the falsehood, the expected accuracy of planning to steadfast is:

(1) $\mathrm{EA}(\mathrm{PL}^{\mathrm{stead}}) = (.5)\mathbf{A}(.99, \mathrm{G}) + (.5)\mathbf{A}(.99, \sim\mathrm{G})$.

Since we are supposing that the relevant impairment won't affect your ability to do what calibrationism recommends, we can expect that the result of planning to calibrate is that you *will* calibrate. This will involve adopting the credence function *.5* both when G is true and when G is false.[16] Thus, the expected accuracy of planning to calibrate is:

(2) $\mathrm{EA}(\mathrm{PL}^{\mathrm{cal}}) = \mathbf{A}(.5, \mathrm{G}) = \mathbf{A}(.5, \sim\mathrm{G})$.

We can rewrite (2) as:

(3) $\mathrm{EA}(\mathrm{PL}^{\mathrm{cal}}) = (.5)\mathbf{A}(.5,\mathrm{G}) + (.5)\mathbf{A}(.5, \sim\mathrm{G})$

We can think of planning to steadfast as being epistemically *risky*: making the plan gives you a 50% chance at a high accuracy score and a 50% chance at a low accuracy score. Planning to calibrate, on the other hand, can be seen as epistemically *conservative.* You get a guaranteed middling level accuracy score. So should we be risky or conservative? We should be conservative, and the argument for this will appeal to the fact that we're using a strictly proper scoring rule. To see why any such scoring rule will favor the conservative plan, it will be helpful to revisit the implications of a scoring rule being strictly proper.

Consider Claire whose credence function is *c* (and recall that this means that Claire assigns credence c to G and credence 1-c to ~G and that we are restricting our attention to Claire's credences in G and ~G). Note that according to the definition of expected accuracy, the expected accuracy that *c* assigns to a credence function *x* is:

---

[16] Of course, you might assign non-zero credence to the proposition that you won't calibrate as a result of so planning. For perhaps you'll forget your plan, get eaten by a lion or go crazy. But since we're comparing steadfasting to calibrating, and we can't expect a significant difference in results in these cases between people who planned to steadfast and people who planned to calibrate, we can safely set them aside.

(4) $EA^c(x) = (c)\mathbf{A}(x, G) + (1\text{-}c)\mathbf{A}(x, \sim G)$

If **A** is strictly proper, it follows that (4) is maximized when *x* assigns credence c to G and credence 1-c to ∼G. That is, Claire thinks that assigning her *own* credences to G and to ∼G will maximize expected accuracy.

Since:

(5) $(.5)\mathbf{A}(x, G) + (.5)\mathbf{A}(x, \sim G)$

is just an instance of (4), it follows that (5) is maximized when *x* assigns a .5 credence to each of G and ∼G – that is, when $x = .5$.

Why is this relevant to the expected accuracy of the calibrationist's and steadfaster's plans? Note that both (1) and (3) are instances of (5). The difference between them is that (1) plugs in *.99* for *x* while (3) plugs in *.5* for *x*. Since (5) is maximized when $x = .5$, (3) must be greater than (1). Thus, the expected accuracy of *planning* to calibrate is greater than the expected accuracy of *planning* to be steadfast.

The calibrationist plan will also do better than planning to assign any other credence to the proposition best supported by the first order evidence, in cases in which the judgments you form on the basis of the first order evidence are only correct 50% of the time. For example, suppose you planned to respond to the hypoxia case by taking a compromise position. Rather than assigning .99 to the proposition that the first order evidence supports, you'll assign, say, a 0.7 credence to the proposition that the first order evidence best supports and a 0.3 credence to its negation. (This kind of position has been called by Tom Kelly "the total evidence" view). Calibrationism will beat any such plan. For you should expect that the result of making this plan in cases of hypoxia will be that 50% of the time you'll have a .7 credence in the truth and 50% of the time you'll have a .7 credence in the falsehood. More generally, the expected accuracy of making a "total evidence" plan will be:

(6) $(.5)\mathbf{A}(k, G) + (.5)\mathbf{A}(k, \sim G)$

where k is the credence that you plan to assign to the proposition that the first order evidence best supports. But, once again, since we're using a strictly proper scoring rule, $k=.5$ is the value that maximizes (6).[17]

In sum, on a plausible view about what would result from planning to be steadfast, *planning* to calibrate does better than *planning* to be steadfast, even though steadfasting does better than calibrating. Planning to calibrate also does better than planning to conform to the alternative views that have been proposed in the literature.

## 6. Conclusion

What the accuracy-based considerations tell us about higher order evidence depends on what quantity we are trying to maximize. On the standard picture, we're trying to maximize the expected accuracy of the credences that would result from *conforming* to some update-procedure. Such a view motivates steadfastness. An alternative picture, according to which we are trying to maximize the expected accuracy of the credences that would result from *planning* to update in a certain way, favors calibrating. So what theory of *rationality* should we accept if the correct theory of rationality is supposed to be, in some sense, accuracy conducive? My own view is that once we've shown that the answer depends on the sense of *accuracy-conducive* one has in mind, there is not really anything of interest left to say concerning which of these notions describes *rationality*. Personally, I like to think of the activity that I and other normative epistemologists are engaged as *deliberating about what to believe*. Epistemologists who see their work as engaging in the what-to-believe project should be more sympathetic to calibrationism since we should be concluding our what-to-believe deliberations with doxastic plans such that *making them* has high expected accuracy. Epistemologists who see what they are doing in other ways may favor steadfastness. For example, an epistemologist might think of

[17] The calibrationist position can also be generalized to cases in which you expect that, due to some cognitive impairment, you'll be *somewhat* unreliable but still, you'll do better than chance. Imagine learning that the chance of a hypoxic pilot arriving at a correct judgment on the basis of her first order evidence is r. What credence should the calibrationist recommend assigning to one's judgment? If you plan to assign credence x to the proposition that you judged on the basis of the first order evidence then you should assign credence r to ending up with credence x to the truth and credence 1-r to ending up with credence x to the falsehood. Thus, the expected accuracy of making such a plan is: $(r)\mathbf{A}(x, G) + (1-r)\mathbf{A}(x, \sim G)$. Since we're using a strictly proper scoring rule, this quantity is maximized when $x=r$. Thus, if you want to maximize expected accuracy, you should plan to assign credence r to the proposition that you judged on the basis of the first order evidence and credence 1-r to its negation.

what she is doing as describing the update procedures that a cognitively unlimited agent *who was certain that she would make no cognitive errors in the future* would endorse, and conform to. On such a view steadfasting will look more attractive.

**References**

Arnold, A. (2013). "Some Evidence is False." *Australasian Journal of Philosophy*. 91 (1): 165-172.

Bronfman, A. (forthcoming). "Conditionalization and not Knowing that one Knows." *Erkenntniss*

Christensen, D. (2010). "Higher Order Evidence." *Philosophy and Phenomenological Research* 81(1):185-215.

Elga, A. (ms.). "Lucky to Be Rational."

Gibbard, A. (2008) "Rational Credence and the Value of Truth." *Oxford Studies In Epistemology Volume 2*. Oxford University Press.

Gibbard, A. (2003). *Thinking How to Live*. Harvard University Press.

Greaves, H. and Wallace, D. (2006). "Justifying conditionalisation: conditionalisation maximizes expected epistemic utility." *Mind* 115(459): 607-632.

Greco, D. and Hedden, B. (ms.)

Joyce, J. (1998). "A Nonpragmatic Vindication of Probabilism." *Philosophy of Science* 65 (4):575-603.

Horowitz, S. (2013). "Immoderately Rational." *Philosophical Studies* 167(1):1-16.

Horowitz, S. and Sliwa, P. (ms.). "Respecting All the Evidence."

Kelly, T. (2010). "Peer Disagreement and Higher Order Evidence." In Alvin Goldman & Dennis Whitcomb (eds.) *Social Epistemology: Essential Readings*. Oxford University Press.

Moss, S. (2011). "Scoring Rules and Epistemic Compromise." *Mind* 120 (480):1053 -1069.

Pettigrew, R. (2012). "Accuracy, Chance, and the Principal Principle." *Philosophical Review* 121 (2):241-275.

Pettigrew, R. (forthcoming)[a]. Accuracy, Risk, and the Principle of Indifference. *Philosophy and Phenomenological Research* 89 (1)

Pettigrew, R. (forthcoming)[b]. *Accuracy and the Laws of Credence*. Oxford University Press.

Rizzierie, A. (2011). "Evidence does Not Equal Knowledge." *Philosophical Studies* 153(2): 235-242.

Schafer, K. (2014). "Doxastic Planning and Epistemic Internalism." *Synthese*

Steel, R. (ms.). "Peer Disagreement, Anticipating Failure, and Avoiding It."

Weatherson, B. (ms.). "Do Judgments Screen Evidence?"