A Note on Unawareness

Jeffrey C. Ely*

May 19, 1998

Abstract

I present a view that justifies dropping the axiom of negative introspection. I then show that the same view justifies dropping the axiom of *AU-Introspection* introduced in Dekel, Lipman, and Rustichini (1998). This view leads naturally to a simple model of knowledge that accommodates non-trivial unawareness. From an analyst's perspective, this model reduces to a standard possibility correspondence framework. Finally, I show that the analogy between negative introspection and AU-introspection is more than suggestive: in any knowledge structure satisfying non-delusion the two axioms are equivalent.

1 Introduction

The famous "curious incident" involving Sherlock Holmes and Watson has been a useful example to illustrate approaches to formalizing "unawareness" or "unforeseen contingencies." It is the leading example in Dekel, Lipman, and Rustichini (1998, hereafter DLR), in which the authors make the point that "standard state-space" models of information and knowledge cannot adequately accommodate a plausible notion of unawareness. In this note, I will revisit this example, provide my own (not terribly novel) perspective on it, and argue that under this perspective (for which I hope to make a compelling case), the DLR critique becomes much easier to bear. I will then suggest a simple model of knowledge which quite naturally allows for unawareness. Of course, given the theorems of DLR this will not be a "standard" model under

*ely@nwu.edu

their definition, but it will be essentially equivalent to such a model, and can be analyzed in the usual way.

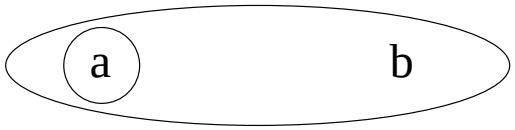

Figure 1: Watson's knowledge structure

Figure 1 illustrates a model of knowledge for Watson. State a is the event in which there was a break-in, state b the event in which there was no break-in.[1] In this non-partitional, but otherwise standard state-space representation, Watson's knowledge structure fails *negative introspection*. At state b, and only at state b, Watson does not know the event a. That is $b = \neg Ka$. But this, together with the fact that at b Watson does not know b, implies $b \in \neg K \neg Ka$. Watson does not know that he does not know a.

To summarize the perspective on this example I will take, the reason we model Watson's knowledge in this way is that it captures the idea that Watson does not fully realize the mechanism which generates his knowledge. In particular, the story goes, Watson would know a if a had occured because he would have been alerted to that fact by a signal. However, being less the sleuth that Holmes is, he does not conceive of the possibility of that signal if he were not to hear it, for example if b occured. Therefore, at b, his limited understanding of the knowledge mechanism provides him with no way of distinguishing a and b.

Formally, when state b has occured, from Watson's perspective, the event "Watson knows a" is empty. From the perspective of the analyst, who has written down this information mechanism, and therefore understands it fully, the event "Watson knows a" is equal to a.

I assert that it is essential that we maintain this distinction. If we were to allow that Watson, at b, understood that $Ka = a$, he would be subject to *reductio ad absurdum*:

1. I don't know that a is true.

[1] In this example, the relevant non-trivial events are singletons, so for notational ease I will avoid brackets.

2. If a were true, I would know it.

It seems crazy to assume that Watson could not deduce from 1 and 2 that a is false and therefore to revise his information model to the one in figure 2. Of course we might explain Watson's information structure using a theory

Figure 2: Holmes' knowledge structure

that implies Watson does not deduce all of the logical implications of his own knowledge. But a model such as that does not seem to be what the curious incident in particular, and unawareness in general are about. Rather, and by the way, simpler from a modeling standpoint, Watson's unawareness of the possibility of a barking dog can be formalized by simply assuming that 2 is not true for Watson at b. Instead, because at b, Watson is unaware of any way of distinguishing a from b, we substitute $Ka = \emptyset$ at the state b.

One advantage of assuming that $Ka = \emptyset$ at b is that it allows a model in which Watson's knowledge is consistent with logical deduction and yet, non-partitional. That is, while the requirement that $Ka = a$ even at b demands that Watson display the above logical inconsistency (a failure of negative introspection), there is nothing in his knowledge at b that is inconsistent with $Ka = \emptyset$. We could lay out for Watson all of the knowledge we attribute to him at b : that both a and b are possible, and that regardless of which has actually occured, he would conclude that both a and b are possible; and this would not lead a logically sophisticated Watson to revise his knowledge.

To summarize, we accept the failure of negative introspection in the example not because we are comfortable with Watson's logical inconsistency, but because we view Watson at b as being unaware of the latent signal, and therefore failing to identify a with Ka. The failure of negative introspection is evident only to the analyst who, unlike Watson at b, sees the full information structure.

2 Unawareness

The purpose of the previous section was to introduce the viewpoint of this note within the context of negative introspection, the failure of which is apparently not controversial. The idea is to sucker the reader into buying into the viewpoint within a neutral context, then to demonstrate that the same viewpoint leads to the main argument of this note: that unwareness can plausibly be captured within an essentially standard possibility correspondence model.

Consider again Watson in state b, but now ask whether Watson is *aware* of the event a. If unawareness is to satisfy the plausibility criterion of DLR, then Watson at b is unaware of a only if at b both $\neg Ka$ and $\neg K\neg Ka$ hold. If we were to require that $Ka = a$, then we would find that $\neg Ka \cap \neg K\neg Ka = b \cap \Omega = b$ and therefore that Watson was unaware of a. However, as DLR point out, it would also be the case that the event $\neg Kb \cap \neg K\neg Kb = \Omega \cap \emptyset = \emptyset$ was not true at b. In words, Watson is aware (i.e. not unaware) of the event b, which happens also to be the event Ua that Watson is unaware of a. Watson's unawareness structure displays a failure of what DLR term AU *introspection*: at b he is aware that he is unaware of a.

The argument in favor of AU-introspection is directly analogous to that in favor of negative introspection: if Watson can identify that it is b he is unaware of, then he must be aware of b. We could directly erase this inconsistency in an *ad hoc* way that parallels the resolution for negative introspection: simply assert that for Watson at b, it is not the case that $b = Ua$. Intuitively, there is nothing in Watson's knowledge at b which should require him to equate b with Ua. It is only the analyst's understanding of the full knowledge mechanism, that allows this inference. Watson at b *is* aware of the event b, which unbeknownst to Watson at b, happens also to be the event "according to the analyst, Watson is unaware of a."

But this *ad hoc* approach is unnecessary because $b \notin Ua$ can in fact be *derived* from the more fundamental postulate from the previous section that $Ka = \emptyset$ at b. Indeed, if $Ka = \emptyset$, then $Ua \subset \neg Ka \cap \neg K\neg Ka = \Omega \cap \neg K\Omega = \emptyset$. In fact, this implication expresses exactly the intuition given in the previous paragraph. Because at b, Watson cannot conceive any way of distinguishing a from b, he equates Ka with $\emptyset$ and from this it follows that $Ua = \emptyset$. Only if Watson at b understood the knowledge mechanism and hence stipulated that $Ka = a$, would he be forced to violate AU-introspection.

3 A Simple Framework

This discussion was intended to motivate a simple approach to capturing unawareness in an essentially standard state-space model. The theme of the approach is that all logical inconsistencies are detectable only by an analyst who can consult the entire knowledge mechanism. At any given state, the decision-maker's view of the world is invulnerable to *reductio ad absurdum.* However, he may maintain different views of the world in different states, and these views may be inconsistent. These inconsistencies survive the decision-maker's introspection because at any given state, he is unaware of whatever mechanism would lead him to a different model at some different state. This sounds a lot like the approach sketched in DLR; certainly the two approaches follow from the same intuition about the nature of unawareness. The present approach however seems simpler and probably more tractable. In fact it can be represented by a standard, non-partitional model.

The knowledge structure is given by a set of states Ω, and a *collection* of partitions $\Pi(\omega)$, one for each $\omega \in \Omega$. The partition $\Pi(\omega)$ represents the decision-maker's model of the world when the state ω is realized. Because it is a partition, the decision maker's model survives negative introspection and AU-introspection at every state. However, from an analyst's perspective, he may violate both. An analyst who knows the mapping $\omega \to \Pi(\omega)$, can reduce the knowledge model into a standard possibility correspondence $\omega \to P(\omega)$, where $P(\omega)$ is the set of states deemed possible by the decision maker according to his model $\Pi(\omega)$ at ω. That is $P(\omega) = \Pi(\omega)(\omega)$. In general, $P(\omega)$ constructed in this way need not be a partition, in which case the analyst will detect a logical inconsistency of which the decision-maker is unaware.

For example, we can model Watson's knowledge and awareness as follows. At a, Watson is aware of the signal, and hence has the discrete partition: $\Pi(a) = \{\{a\}, \{b\}\}$. At b, Watson is unaware of the signal and hence has the trivial partition $\Pi(b) = \{\{a, b\}\}$. We can derive Watson's (now state-dependent) knowledge operator: $K_\omega : \Omega \to 2^\Omega$. At b, according to the model of the world that Watson holds at b, the set of states at which Watson knows that a is true, $K_b(a)$ is empty. This matches the postulate from the informal argument above. For completeness, $K_b(b) = \emptyset, K_a(a) = a, K_a(b) = b$. The analyst reduces this knowledge structure to the usual non-partitional possibility correspondence in the figure. Dropping the subscripts to denote the analyst's model of Watson's knowledge, we have $K(a) = a, K(b) = \emptyset$ as desired.

4 Equivalence of Introspection

I presented an informal argument that suggests that a willingness to drop the requirement of negative introspection, leading to non-partitional information, should imply a willingness to drop AU-introspection. I now make this argument formal by showing that these axioms are equivalent in any knowledge structure which satisfies non-delusion.

A *knowledge structure* is a pair (Ω, P) where Ω is a set of states of the world and P is a correspondence from Ω into itself. From any knowledge structure, the corresponding knowledge operator K can be derived. The following fact about knowledge operators will be used below.

$$\omega \in \neg KF \Rightarrow P(\omega) \cap \neg F \neq \emptyset \qquad (*)$$

The knowledge structure satisfies *non-delusion* if $KE \subset E$ for every event E under the derived knowledge operator. The *unawareness operator* derived from K is the correspondence U from 2^Ω into itself such that $UE = \neg KE \cap \neg K\neg KE$ for every event E. [2] The result is about the equivalence of the following two axioms on knowledge and unawareness:

Negative Introspection $\neg KE \subset K\neg KE$

AU Introspection $UE \subset UUE$.

Proposition 1 *Suppose (Ω, P) satisfies non-delusion and let K and U be derived knowledge and unawareness operators, respectively. Then U satisfies AU-introspection if and only if K satisfies negative introspection.*

Proof: First suppose K satisfies negative introspection. Note that negative introspection is equivalent to $\neg KE \cap \neg K\neg KE = \emptyset$ for every event E. But this implies $UE = \emptyset$ for every E and therefore that AU-introspection is trivially satisfied.

Now we show that if K fails negative introspection then U fails AU-introspection. Failure of negative introspection implies the existence of an event E such that $UE = \neg KE \cap \neg K\neg KE \neq \emptyset$. AU introspection requires $UE \subset UUE$. By the definition of an unawareness operator, the latter set is contained in $\neg KUE \cap \neg K\neg KUE$. Thus, AU-introspection requires both

[2] DLRdefine a *plausible* unawareness operator to be any U for which $UE \subset \neg KE \cap \neg K\neg KE$. The proposition below will hold for any operator U such that $UE \neq \emptyset$ for at least one E such that $\neg KE \cap \neg K\neg KE \neq \emptyset$. Any plausible U which does not satisfy this condition is trivial.

1. $UE \subset \neg KUE$

2. $UE \subset \neg K \neg KUE$

Let $\omega \in UE$ By 2, AU introspection requires $\omega \in \neg K \neg KUE$. By $(*)$, this implies $P(\omega) \cap KUE \neq \emptyset$. Thus KUE is not empty and by non-delusion is contained in UE. But this contradicts 1. ■

The proposition tells us that if we are willing to drop negative introspection (a must if we are to allow for unawareness) this entails the willingness to accept awareness of unawareness. DLR view this as undesirable. But the main point of this note is the discussion in Sections 1 through 3, intended to argue that this shouldn't be a matter for concern.

References

DEKEL, E., B. L. LIPMAN, AND A. RUSTICHINI (1998): "Standard State-Space Models Preclude Unawareness," *Econometrica*, 66(1), 159–173.