AN ILLOCUTIONARY LOGICAL EXPLANATION OF THE SURPRISE EXECUTION

1. THE PROBLEM The surprise execution or surprise examination has been much discussed in the philosophical literature. Two recent discussions are due to faculty members in my own university--these are *Thalos 1997* and *Shapiro 1998*. Since attempts to resolve/explain the alleged paradox have now become a cottage industry at SUNY Buffalo, I have decided to try my hand at a solution. I know of no previous attempts that get it entirely right, but I think the resources of illocutionary logic provide what is needed to clear up the matter.

Let me set up the problem as follows. On Friday, our prisoner is found guilty of a terrible crime, and sentenced to be executed at noon on one day in the following week. It is part of the punishment that the prisoner will not know beforehand the day of his execution--he will be taken by surprise at noon on whatever day is selected.

The problem is that the prisoner knows the verdict, and he knows that the execution is supposed to come as a surprise. This gives him grounds for eliminating Saturday as the day of his execution, for the execution would be no surprise on that day. Once Saturday has been eliminated as a possible day for execution, he can use a similar argument to eliminate Friday. In fact, he can keep going and eliminate every day as a day when he can be executed by surprise. If being surprised were a necessary condition for being executed, he might hope to avoid the execution altogether.

We may suppose that by good reasoning the prisoner eliminates every day of the week as a day for his execution. Still, when they come for him on Tuesday, he is surprised.

The puzzle here is to understand and explain how the prisoner's reasoning has led him astray, or, at least, how it has let him incorrectly to think that the sentence cannot be carried out. Has correct reasoning from true premisses led to a false conclusion?

2. OTHER ATTEMPTS I am not going to survey the entire literature on this puzzle, but I will comment briefly on *Thalos 1997* and *Shapiro 1998*. In *Thalos 1997*, the puzzling aspect of the surprise execution is diagnosed as a matter of conflicting aims on the part of the judge and the prisoner. The judge intends to have the prisoner executed by surprise. The prisoner is trying not to be taken by surprise. The judge deliberately creates the conflict by her sentence, and the prisoner cannot by his reasoning overcome the conflict. Thalos' diagnosis isn't wrong, but it doesn't give us everything we would like. The prisoner reasons correctly but reaches an incorrect conclusion. However, his premisses seem to be correct (with respect to the story). The account to follow provides the conceptual resources that are needed to understand and characterize the prisoner's beliefs and his reasoning. This will enable us to determine that the prisoner's premisses are responsible for his mistaken conclusion. His reasoning is blameless.

In *Shapiro 1998*, the paradox is approached in terms of procedures rather than inferences or arguments. The prisoner needs a procedure he can use for determining on which days he can legitimately be executed. When the prisoner reached the conclusion that no day is a day when he can be executed by surprise, the prisoner was using an inappropriate serial procedure. If he had chosen an appropriate serial procedure or an appropriate parallel procedure, he would have realized that many days are possibilities for a surprise execution. Shapiro’s discussion is clear, entertaining, and ingenious. However, it doesn’t give us what we want: an explanation of how the prisoner’s argument has gone wrong. In deductively correct reasoning from a set of beliefs to a conclusion, it doesn’t make any difference which beliefs are considered first. Any one chain of reasoning from premisses in the set is independent of other such chains. Shapiro has confused arguments with algorithms, and has failed to give an analysis of what is wrong with the prisoner’s argument. We want to understand how it is that by reasoning correctly from premisses which he is justified in accepting (which are, in fact, true), the prisoner arrives at conclusions which are incorrect.

3. LINGUISTIC ACTS In *Kearns 1997*, I present a basic system of propositional illocutionary logic. In this paper, I will just sketch this system and say enough about it to make clear how it applies to the present situation. I won’t go through all the details of the semantic account, or prove various kinds of adequacy for the system I develop in this paper.

It is important to understand that illocutionary logic must be considered in a speech act framework. From this perspective, the basic linguistic reality is constituted by acts of using expressions--by speech acts or linguistic acts, and by skills and dispositions for performing linguistic acts. The expression ‘speech act’ suggests an act where someone speaks aloud. But I don’t intend this. The person who speaks or writes or thinks with words is performing speech acts/linguistic acts. The term ‘linguistic act’ is less likely to be misleading than ‘speech act,’ but I will use the two expressions interchangeably. It isn’t only the person who produces expressions who performs speech acts; so does the person who reads or who listens with understanding.

Expressions used to perform linguistic acts are not themselves meaningful. The primary bearers of meaning are linguistic acts. Expressions are the bearers of syntactic features, and can even be regarded as syntactic objects. Certain expressions are conventionally used to perform certain meaningful acts. The meanings commonly attributed to expressions are the meanings of acts the expressions are conventionally used to perform. However, the primary source of meaning is the language user’s intentions, and not convention. I can by mistake use the wrong expression to refer to someone, but I refer to the person I intend. (When this happens, my audience may fail to know which person I intend.) Still, in using language it is customary to intend the meanings with which the words are conventionally associated.

Different kinds of expressions are used to perform different kinds of acts. Names and descriptions can be used to perform referring acts. Predicates can be used to perform various sorts of predicative acts. Whole sentences are used to perform *sentential acts*. A *propositional act* is a sentential act that can appropriately be evaluated in terms of truth and falsity. A propositional act can be accepted (as true) or rejected (as false). A propositional act can also be supposed true, to determine what follows from it. I understand an *assertion* to be a propositional act which is performed and accepted all at once. This understanding is at odds with accepted terminology, for an assertion in my sense does not require an audience. And on my usage, there is no such thing as an insincere assertion. (On conventional usage, a sincere assertion is one in which the propositional act is accepted.) What others call an insincere assertion, I must regard as a pretended assertion.

An assertion is an *illocutionary act*--it is constituted by a propositional act performed with a certain illocutionary *force*. The same (kind of) propositional act can be performed on another occasion with a different illocutionary force, or with no illocutionary force. The word ‘statement’ is often used as a near synonym for ‘assertion,’ but I will use it instead for propositional acts. (The word ‘statement’ is less cumbersome than ‘propositional act.’) So on my usage, some statements are asserted and some aren’t. Propositional acts or statements are as close as I come to abstract true or false propositions. But a propositional act exists only if and when it is performed. However, we can represent propositional acts that no one performs, just as we can draw pictures of events that never occur.

My understanding of linguistic reality makes some problems for artificial logical languages. Nobody speaks or writes or thinks with these languages. So in an important sense, these aren’t really languages, though this won’t keep me from speaking of logical languages. The expressions of an artificial logical language are not regarded as candidates for being used to perform linguistic acts. Instead the expressions are *representations* of linguistic acts performed with expressions of natural languages. Truth conditions of sentences in logical languages are really truth conditions of statements represented by these sentences, for it is speech acts, not sentences, which are true or false. Logical deductive systems simply codify those artificial-language expressions which represent logically true statements or logically valid argument sequences.

4. ILLOCUTIONARY LOGIC The language *L* contains atomic sentences, sentences formed with these connectives:

~, v, &, ⊃

and sentences formed with operators to be explained below. The atomic sentences and sentences formed from atomic sentences with connectives and operators are *plain* sentences of *L*.

It is a feature of my systems of illocutionary logic that the artificial languages contain expressions for indicating/representing illocutionary force. The symbol introduced by Frege is used for assertion/acceptance: ⊢. Rotated 180°, this becomes the symbol for rejection: ⊣. Someone who rejects a statement takes the statement to be false. To suppose a sentence true, I use the top half of the assertion sign: ⎵. Rotated 180°, this becomes the symbol for supposing a sentence false: ¬. Each illocutionary force indicator marks a force with which a statement can be performed. Declining to perform one of these acts is also an illocutionary act. Someone who declines to accept a statement is not taking the statement to be false--he isn’t taking any stand about the truth or falsity of the statement. The following are symbols for declining to perform the various acts: *x*⊢, *x*⊣, *x*⎵, *x*¬.

In *L*, only two illocutionary-force indicators will be employed, ‘⊢’ and ‘⎵.’ I am making this restriction in order to keep *L* manageably simple. If *A* is a plain sentence of *L*, then ⊢*A* and ⎵*A* are *completed* sentences of *L*; there are no other completed sentences.

The illocutionary force indicators cannot be iterated, so we have no sentences like these: $\vdash\vdash A$, $\llcorner\vdash A$. And a completed sentence cannot be a component of a larger sentence. The following are not allowed (they make no sense): $[\vdash A \,\&\, \llcorner B]$, $\sim\vdash A$.

The deductive system S is a natural deduction system which employs tree proofs. Only completed sentences occur as steps in these tree proofs. The rules for constructing tree proofs take account of illocutionary force. Some *elementary* rules are illustrated below.

& Introduction   *& Elimination*

$$\frac{?A \quad ?B}{?[A \,\&\, B]} \qquad\qquad \frac{?[A \,\&\, B]}{?A} \qquad \frac{?[A \,\&\, B]}{?B}$$

Each question mark is either '$\vdash$' or '$\llcorner$.' If one premiss has the force '$\llcorner$,' then so does the conclusion. Otherwise, the conclusion is an assertion.

The following are instances of *& Introduction*:

$$\frac{\vdash A \quad \vdash B}{\vdash [A \,\&\, B]} \qquad\qquad \frac{\llcorner A \quad \vdash B}{\llcorner [A \,\&\, B]}$$

From the second example, we see that the conclusion as well as the premiss of an argument can have the force of a supposition.

$\supset$ Elimination (Modus Ponens)

$$\frac{?A \quad ?[A \supset B]}{?B}$$

The conditions on this rule are the same.

The top steps (the beginning steps) of a proof are its *initial* steps. An initial supposition of a proof is a *hypothesis* of the proof. An initial assertion is not a hypothesis. Proofs in S represent arguments made with the illocutionary acts represented by the statements in the proof. Any statement can serve as a hypothesis in a proof, but it is only appropriate for a person to begin from initial assertions that represent statements which she accepts.

If $A_1,\ldots, A_n, B$ are (plain) sentences of L, then $A_1,\ldots, A_n / B$ is an *argument sequence*; $A_1,\ldots, A_n$ are the premisses and B is the conclusion.

A proof in S from initial assertions $\vdash A_1,\ldots, \vdash A_m$ and initial suppositions $\llcorner B_1,\ldots, \llcorner B_n$ to conclusion $\llcorner C$ establishes '$B_1,\ldots, B_n / C$' as a result. This result is relative to knowledge/belief that includes $A_1,\ldots, A_m$. If the conclusion is $\vdash C$, then $n = 0$. Such a proof establishes C as a result with respect to knowledge/belief that includes $A_1,\ldots, A_m$.

This proof:

```
 ⌞A      ⊢B
------------ &I
⌞[A & B]        ⊢[[A & B] ⊃ C]
----------------------------------- ⊃E
                 ⌞C
```

establishes ‘*A / C*’ as a result relative to knowledge/belief that includes *B, [[A & B] ⊃ C]*. And this proof:

```
⌞[A & B]                ⌞[A & B]
----------- &E          ----------- &E
     ⌞B                       ⌞A
     ------------------------------ &I
              ⌞[B & A]
```

establishes ‘*[A & B] / [B & A]*.’

The non-elementary rules of *S cancel* or *discharge* hypotheses. In illustrating non-elementary rules, I use braces to enclose the hypotheses that are cancelled:

⊃ Introduction

```
  {⌞A}      If the only uncancelled hypotheses in the (sub-)proof leading to
            the sentence on the line are those in braces, then the conclusion is an
   ⌞B       assertion. Otherwise the conclusion is a supposition.
------------
?[A ⊃ B]
```

~ Elimination

```
{⌞~A}                    {⌞~A}         {⌞~A}      {⌞~A}      The condition is
                                                             the same.
 ⌞B       ?~B         ?B       ⌞~B      ⌞B          ⌞~B
----------------      ---------------   --------------------
      ?A                    ?A                ?A
```

The proof below justifies the assertion of a form of the Law of Exportation:

```
 x       x
⌞A      ⌞B
------------ &I                 x
⌞[A & B]               ⌞[[A & B] ⊃ C]
-------------------------------------------- ⊃E
                ⌞C
            ----------- ⊃I, drop ‘⌞B’
              ⌞[B ⊃ C]
        -------------------- ⊃I, drop ‘⌞A’
        ⌞[A ⊃ [B ⊃ C]]
------------------------------------------ ⊃I, drop ‘⌞[[A & B] ⊃ C]’
⊢[[[A & B] ⊃ C] ⊃ [A ⊃ [B ⊃ C]]]
```

An ‘*x*’ is placed above occurrences of hypotheses that are cancelled. Since all hypotheses in the proof are cancelled, the proof establishes the correctness of the conclusion, and the conclusion is asserted.

5. COMMITMENT-BASED SEMANTIC CONCEPTS An actual argument/proof is very much a first-person affair. The statements of the argument are either asserted or supposed by the arguer. Her assertions are statements that she accepts. Someone else can appraise her argument, but if the initial assertions are not accepted by the someone else, then the argument is not one that the someone else can legitimately make. An argument with explicit illocutionary force markers must be evaluated with respect to individual people, and particular times, for an argument might be correct/appropriate for one person but not for someone else. It might also be incorrect/inappropriate for a given person at one time, but be correct/appropriate for that person at a later time.

In the Twentieth Century, at least since the work of Tarski, semantic investigations have focused exclusively on truth conditions. But semantic concepts based on truth conditions are not the only ones that can be rigorously defined and systematically explored. Illocutionary logic supplements the study of truth-conditional concepts with the study of concepts based on *rational commitment*.

The word ‘commitment’ is sometimes used to talk about obligation, especially moral obligation. For example, someone might say that two people make a commitment to one another in marriage. I have added the qualifying adjective ‘rational’ to distance the commitment I have in mind from moral obligation/moral commitment. Rational commitment can be generated by a decision to perform an action. If I decide to stop at the bank to get money on my way home from work, I am committed to doing this. But I have no obligation--I am not “supposed” to stop on the way home. I can cancel the commitment by changing my mind. I can also forget, and drive straight home without going to the bank. I will have failed to honor my commitment, and I may regret this, but I will not have done anything immoral or sinful.

Deciding to act generates a commitment to acting. Doing one thing can also generate a commitment to do something else. Putting bananas in my cart at the supermarket commits me to putting them on the conveyer belt in the checkout line. Some rational commitments are absolute, come-what-may commitments, like my commitment to stop at the bank. Some are conditional, and only take effect when the condition is satisfied, like my commitment to shut the upstairs windows if it rains while I am at home.

Accepting some statements rationally commits me to accept others and to reject still others (so long as I continue to accept the original statements). If I accept the statement that Peter is presently in his office at school, then I am committed to reject the statement that Peter is now at home, and to accept the statement that he is currently on campus. The commitment to accept a statement is conditional--I am committed if the question comes up and I choose to think about it. And the commitment can be cancelled; if I discover that Peter isn't in his office, I will no longer accept the statement that he is, and I will no longer be committed to the other consequences.

The truth conditions of a statement/propositional act are independent of illocutionary force. So are truth-conditional semantic concepts. For example, statements $A_1,..., A_n$ *truth-conditionally entail* statement B if, and only if (from now on: iff), there is no way to satisfy the truth conditions of $A_1,..., A_n$ without satisfying those of B.

We can define concepts based on commitment, which are not independent of illocutionary force. Statements $A_1,..., A_n$ *basically entail* B iff accepting $A_1,..., A_n$ (rationally) commits a person to accepting B. An argument sequence is *basically valid* iff its premisses basically entail its conclusion. Statements $A_1,..., A_n$ *suppositionally entail* B iff supposing $A_1,..., A_n$ commits a person to supposing B. We can also consider commitments generated by combinations of assertions and suppositions.

Truth-conditional and basic entailment coincide for the most part, but not entirely. For an argument sequence '*A / I believe that A*' is basically valid without being truth-conditionally valid. The argument sequence also fails to be suppositionally valid, for supposing a statement is supposing it to be true, not supposing it to be known/believed.

Forms of entailment are *general* semantic features; they are not associated with a particular class of expressions or linguistic forms. Logical special cases of general features are associated with logical forms in artificial logical languages. I use the word '*implication*' for the logical special cases of entailment.

Commitment-based concepts are important with respect to the human practice of giving arguments. It is the recognition of commitment that propels a person from the premisses to the conclusion of an argument. That the premisses of an argument truth-conditionally entail the conclusion won't make the argument effective or convincing. The arguer or her audience needs to *recognize* the truth-conditional entailment--once they do this, they have recognized a commitment to accept the conclusion once they accept the premisses.

6. MOORE'S PARADOX In *Kearns 1997*, artificial language $L_{.75}$ is provided with both a truth-conditional semantics and a commitment-based semantics, and the relations between the two semantic

accounts are explored. Here I am considering the language *L*, but I will develop the two kinds of semantic account for a fragment of *L*. This will be sufficient to show how the semantics works. I won't prove results about the fragment and its semantics. We don't need more details, or proofs, in order to understand what is going on in the prisoner's argument about his execution.

With respect to statements and illocutionary acts, various kinds of conflict (and its absence) can arise. Conversationally, the words 'consistent' and 'inconsistent' are used for semantic features; I will use the words in this way. One or more statements are *inconsistent*, and two or more statements are *incompatible* iff it isn't possible for their truth conditions to be simultaneously satisfied. If it is possible to satisfy the truth conditions, there is consistency or compatibility. Consistency and inconsistency apply to propositional acts independently of illocutionary force. If someone *accepts* inconsistent statements, I will describe her acts as *incoherent*. Incoherence characterizes assertions, denials, and beliefs, but not (merely) their propositional content. A person whose beliefs commit her to accepting incoherent statements has incoherent beliefs. The fundamental form of incoherence is probably performing and declining to perform the same linguistic act (without changing one's mind in between).

Truth conditions of sentences in artificial languages and the statements these represent are independent of the person who uses the sentences or makes the statements. It is somewhat different with commitment conditions. In providing a semantic account which assigns values indicating acceptance (+), rejection (-), or neither one (*n*), we regard the assignment as reflecting the linguistic acts of some person, whom I call the *designated subject*. Given such an assignment, arguments/proofs are then evaluated from the standpoint of the designated subject.

Truth-conditional semantic accounts for artificial languages treat sentences and the statements they represent as highly specific. The statements concern particular situations or states of affairs, and have definite truth values. While we can consider commitment conditions of highly specific statements, we can also consider less specific statements.

The statements that can be made with this sentence:

> Today is Tuesday.

range over all the days when there are people around who speak English. Some of these statements are true and others false. Different true statements can be concerned with different Tuesdays. If we consider all of these statements as forming a single kind of statement, we have a highly general statement kind, not a highly specific one. In contrast, this sentence:

> June 17, 1998 was a Tuesday.

determines a highly specific statement kind and concerns a particular state of affairs. This statement kind still has multiple instances, but the instances will all have the same truth value.

In exploring the commitments of illocutionary acts, we are really investigating kinds of illocutionary acts. *The* statement that June 17, 1998 was a Tuesday is a single *kind* of statement but not a single statement, for each time a person says this, a different instance of the kind is produced. In providing a semantic account, we shall limit our attention to highly specific statement kinds describing particular situations.

We shall focus on explicit beliefs understood this way: A person who performs and accepts propositional act *A* has an *explicit* belief at that time *that A*. Accepting a statement is to be understood as accepting it now (at the moment of utterance) and for the future. A similar remark applies to an act of rejecting a statement; in rejecting a statement as false, we commit ourselves to continue to reject it in the future. The person who performs and accepts *A* continues to have an explicit belief that *A* until she either changes her mind about *A* (or about the reasons that led her to accept *A*) or she simply forgets that she has accepted *A* and that she remains committed to accepting *A*.

A statement '*I believe that A*' made at different moments is indexical to its moment of utterance. We will not consider all statements made with '*I believe that A*' (at different times) to belong to or constitute a single kind. There is a different statement kind '*I believe that A*' for each moment, and the semantics will be developed for one such kind. A *moment* is determined by the set of explicit beliefs and the set of explicit disbeliefs of the designated subject at some time. Given the explicit beliefs and disbeliefs, there is a fixed set of statements that the designated subject is committed to accept, and a fixed set he is committed to reject. So long as the designated subject accepts only statements he is committed to accept and rejects only statements he is committed to reject, he stays in the same moment. Once he forms a new belief or disbelief which generates new commitments, the previous moment is ended.

If the designated subject accepts statement *A* at a given moment, he is committed at that moment to accept '*I believe that A*.' He will be committed at all later moments to accept it. But at later moments, he could not use '*I believe that A*' to accept the earlier modal statement, for the later statement will be indexical to its moment of utterance. We are not providing L_f with the resources to specify earlier moments of utterance.

Moore's Paradox concerns statements made with sentences of the form '*A but I don't believe that A*.' His particular example was 'It's raining but I don't believe it.' Using the resources of *L* we can represent statements involving the same idea like this: *[A & ~I believe that A]*. Moore recognized that the statements are truth-conditionally consistent, but he also realized that there is something wrong with accepting these statements, and he wanted to characterize just what is wrong. In *Moore 1944*, he indicates that there is a sense of 'imply' according to which the person who says *A* (It's raining) implies that she believes *A*. In *Austin 1962*, the author contrasts this sort of implication with what he calls entailment, which is probably truth-conditional entailment. The implication in question is nothing but basic entailment. Austin seems to have thought that basic entailment and truth-conditional entailment are disjoint from one another, but the two actually coincide to a large extent.

There is really nothing paradoxical about Moore's Paradox. A propositional act '*A & ~I believe that A*' is (truth-conditionally) consistent, but it is incoherent for a person to accept this statement. The person who accepts *A* is committed to accept '*I believe that A*.' If she also accepts '*~I believe that A*' instead, then she is committed to accept inconsistent statements. However, while the idea of what is going on is simple, it isn't a simple task to devise a semantic account which accommodates and reconciles truth conditions and commitment conditions for a language containing an '*I believe that*' operator. I will carry out this somewhat complicated task below, to show that our understanding and resolution of Moore's Paradox are acceptable.

The fragment L_f contains plain atomic sentences, the connectives listed earlier, the (modal) operator '*I believe that*,' and plain sentences composed from them. It also contains completed sentences ⊢*A* and ⌞*A*.

An *interpreting function of* L_f is a function which assigns one of T, F to each atomic sentence of L_f. An interpreting function determines the truth values of all (plain) sentences of L_f.

The commitment values of sentences of L_f are + (for sentences representing statements that the designated subject accepts or is committed to accept), - (for sentences representing statements he is committed to reject), and *n*. The following matrices partly characterize the commitment "content" of the connectives:

A	*~A*
+	-
n	*n*
-	+

A	*B*	*[A v B]*	*[A & B]*	*[A ⊃ B]*
+	+	+	+	+
+	*n*	+	*n*	*n*
+	-	+	-	-
n	+	+	*n*	+
n	*n*	*n, +*	*n, -*	*n, +*
n	-	*n*	-	*n*
-	+	+	-	+
-	*n*	*n*	-	+
-	-	-	-	+

A matrix line like this:

A	*B*	*[A v B]*
+	*n*	+

does not indicate that if *A* is (explicitly) believed and *B* is neither believed nor disbelieved, then '*[A v B]*' is (explicitly) believed. The matrices are for *commitment.* The line indicates that if the designated subject either accepts or is committed to accept *A*, and is committed in neither direction toward *B*, then he is committed to accept '*[A v B]*.' In some cases, the commitment values of component statements are insufficient to (completely) determine the values of compound statements. The commitment matrices must be supplemented to provide an adequate commitment semantics.

A *commitment valuation* is a function which assigns (exactly) one of +*, n,* - to each sentence of L_f. Not all commitment valuations are intuitively satisfactory, so we must narrow the class of commitment valuations to obtain those that *are* satisfactory. In doing this we will relate commitment valuations to truth condition valuations.

The sentences of L_f which do not contain '*I believe that*' are conceived as having truth conditions independent of the designated subject's beliefs and disbeliefs. Truth conditions of sentences containing '*I believe that*' may be determined in part by the commitment semantics. In developing this semantics, we restrict our attention to commitment valuations that are so related to interpreting functions that the designated subject's beliefs might all be true. This means that the designated subject's beliefs are coherent; they will be true just in case they are based on the interpreting function which reflects the actual state of affairs. This idealization enables us to identify the arguments that are correct from the standpoint of a person whose beliefs are true, since we all think, of each of our beliefs, that it is true.

Given an interpreting function *f*, the truth values of those sentences that don't contain the operator '*I believe that*' are determined independently of what the designated subject believes or disbelieves. We conceive the designated subject to start with explicit beliefs and disbeliefs concerning sentences without the operator '*I believe that*.' These starting beliefs and disbeliefs will then completely determine which '*I believe that A*' sentences the subject is committed to accept and reject. A logically correct designated subject will believe/accept such a sentence only if he is committed to accept it and he reaches it by deductively correct reasoning. He will reject such a sentence only if he is committed to regard it as false and establishes this by reasoning. There will be many false '*I believe that A*' sentences which are not rejected. The designated subject may be uncertain whether or not he is committed to accept *A*; in that case, he will also be uncertain about '*I believe that A*.'

A commitment matrix is needed for the operator '*I believe that*.' We will use this:

A	*I believe that A*
+	+
n	*n,* -
–	-

A sentence '*[B & ~I believe that B]*' might have value T, but it can never coherently have value +. If it has value T, it can't have value -, for this value is for statements that are to be rejected as false. For a subject whose beliefs are true, a true statement '*[B & ~I believe that B]*' will have commitment value *n*. However, '*I believe that [B & ~I believe that B]*' will have value -.

Let $\mathscr{E}_1$, $\mathscr{E}_2$ be commitment valuations. *$\mathscr{E}_2$ extends $\mathscr{E}_1$* iff for every plain sentence A of L_f, if $\mathscr{E}_1(A) = +$, then $\mathscr{E}_2(A) = +$, and if $\mathscr{E}_1(A) = -$, then $\mathscr{E}_2(A) = -$.

To develop the commitment semantics, it is convenient to decompose L_f into a sequence of languages L_{1f}, L_{2f}... The language L_{1f} contains all the plain sentences of L_f that don't contain occurrences of the operator '*I believe that.*' There are no other (plain) sentences in L_{1f}.

Given L_{mf}, the (plain) sentences of $L_{(m+1)f}$ are obtained as follows:
(1) The sentences of L_{mf} are sentences of $L_{(m+1)f}$;
(2) If A is a sentence of L_{mf}, then '*I believe that A*' is a sentence of $L_{(m+1)f}$;
(3) If A, B are sentences of $L_{(m+1)f}$, then so are $\sim A$, $[A \vee B]$, $[A \& B]$, $[A \supset B]$.

Let f be an interpreting function of L_f. The *(truth-condition) valuation of L_{1f} determined by* f assigns f's values to atomic sentences, and assigns truth-table values to compound sentences.

Let f be an interpreting function of L_f. Then a commitment valuation $\mathscr{E}$ of L_{1f} is *based on f* iff $\mathscr{E}$ assigns + only to sentences that are true for the valuation determined by f, and assigns - only to sentences that are false for the valuation determined by f. If $\mathscr{E}$ is based on f, not all true sentences need to be assigned +, but all the sentences assigned + will be true (for the valuation determined by f).

A commitment valuation $\mathscr{E}$ (of L_f or L_{1f}, L_{2f}, etc.) is a *minimally acceptable valuation* iff $\mathscr{E}$ agrees with the matrices given earlier. If f is an interpreting function of L_f and $\mathscr{E}$ a minimally acceptable valuation of L_{1f} that is based on f, then $<f, \mathscr{E}>$ is a *minimally acceptable pair for L_{1f}*.

Let f be an interpreting function of L_f. Let $\mathscr{E}_0$ be a commitment valuation of L_f which assigns + and - only to sentences of L_{1f} (so sentences of L_f that aren't in L_{1f} are assigned n), and whose restriction to sentences of L_{1f} is based on f. Then $\mathscr{E}_0$ is an *initial commitment valuation of L_f* and *is based on f*. We will treat an initial commitment valuation as a commitment valuation of L_{1f} instead of introducing an additional symbol for its restriction to sentences of L_{1f}. Initial commitment valuations reflect the designated subject's "starting" beliefs at a given moment.

Let f be an interpreting function of L_f and let $\mathscr{E}_0$ be an initial commitment valuation based on f. The *first commitment valuation determined by $\mathscr{E}_0$* is the function $\mathscr{E}_1$ such that for every (plain) sentence A of L_{1f},

(1) If for every minimally acceptable pair $<f^*, \mathscr{E}^*>$ such that $\mathscr{E}^*$ is an extension of $\mathscr{E}_0$, $f^*(A) = \mathrm{T}$, then $\mathscr{E}_1(A) = +$;

(2) If for every minimally acceptable pair $<f^*, \mathscr{E}^*>$ such that $\mathscr{E}^*$ is an extension of $\mathscr{E}_0$, $f^*(A) = \mathrm{F}$, then $\mathscr{E}_1(A) = -$;

(3) Otherwise, $\mathscr{E}_1(A) = n$.

The first commitment valuation assigns + to the sentences that the designated subject is committed to accept by the beliefs and disbeliefs reflected by $\mathscr{E}_0$, and assigns - to the sentences he is committed to reject.

Let f be an interpreting function of L_f. Let $\mathscr{E}_0$ be an initial commitment valuation based on f, and let $\mathscr{E}_1$ be the first commitment valuation determined by $\mathscr{E}_0$. Then the (*truth condition*) *valuation of* L_{2f} *determined by* $<f, \mathscr{E}_1>$ is as follows:

(1) If A is a sentence of L_{1f}, then $<f, \mathscr{E}_1>(A) = f(A)$;

(2) If A is a sentence of L_{1f}, then $<f, \mathscr{E}_1>$*(I believe that A)* = T iff $\mathscr{E}_1(A) = +$;
$<f, \mathscr{E}_1>$*(I believe that A)* = F otherwise;

(3) The remaining sentences have truth-table values.

Let f be an interpreting function of L_f and let $\mathscr{E}_0$ be an initial commitment valuation based on f. Let $\mathscr{E}_1$ be the first commitment valuation determined by $\mathscr{E}_0$. A commitment valuation $\mathscr{E}$ of L_{2f} is *based on* $<f, \mathscr{E}_1>$ iff $\mathscr{E}$ assigns + only to sentences true for the valuation determined by $<f, \mathscr{E}_1>$, and assigns - only to sentences false for this valuation.

Let f be an interpreting function and $\mathscr{E}_0$ be an initial commitment valuation based on f. Let $\mathscr{E}_1$ be the first commitment valuation determined by $\mathscr{E}_0$. And let $\mathscr{E}$ be a minimally acceptable valuation of L_{2f} that is based on $<f, \mathscr{E}_1>$. Then $<<f, \mathscr{E}_1>, \mathscr{E}>$ is a *minimally acceptable pair for* L_{2f}.

Let f be an interpreting function of L_f and let $\mathscr{E}_0$ be an initial commitment valuation based on f. Let $\mathscr{E}_1$ be the first commitment valuation determined by $\mathscr{E}_0$. The *second commitment valuation determined by* $\mathscr{E}_0$ is the function $\mathscr{E}_2$ such that for every (plain) sentence A of L_{2f},

(1) If for every minimally acceptable pair $<<f^*, \mathscr{F}_1>, \mathscr{F}>$ such that $\mathscr{F}$ extends $\mathscr{E}_1$ (for the sentences of L_{1f}), $<f^*, \mathscr{F}_1>(A) = \mathrm{T}$, then $\mathscr{E}_2(A) = +$;

(2) If for every minimally acceptable pair $<<f^*, \mathscr{F}_1>, \mathscr{F}>$ such that $\mathscr{F}$ extends $\mathscr{E}_1$ (for the sentences of L_{1f}), $<f^*, \mathscr{F}_1>(A) = \mathrm{F}$, then $\mathscr{E}_2(A) = -$;

(3) Otherwise, $\mathscr{E}_2(A) = n$.

This process is continued forever, yielding (truth condition) valuations of L_{1f}, L_{2f}... and commitment valuations $\mathscr{E}_1, \mathscr{E}_2$,... determined by $\mathscr{E}_0$.

Given f, $\mathscr{E}_0$ as above, the *final commitment valuation determined by* $\mathscr{E}_0$ (or, simply, *the* commitment valuation determined by $\mathscr{E}_0$) is the function $\mathscr{E}$ such that for every sentence A of L_f,

(1) If for some $\mathscr{E}_m$ determined by $\mathscr{E}_0$, $\mathscr{E}_m(A) = +$, then $\mathscr{E}(A) = +$;

(2) If for some $\mathscr{E}_m$ determined by $\mathscr{E}_0$, $\mathscr{E}_m(A) = -$, then $\mathscr{E}(A) = -$;

(3) Otherwise, $\mathscr{E}(A) = n$.

Given $f, \mathscr{E}_0$ as above, each $\mathscr{E}_{m+1}$ extends $\mathscr{E}_m$, and each $\mathscr{E}_m$ is based either on f (for $m = 1$) or on $<f, \mathscr{E}_{m-1}>$. And $\mathscr{E}$ is a minimally acceptable valuation of L_f that extends each $\mathscr{E}_m$ for $m \geq 0$, and is such that if $\mathscr{E}(A) = +$, then for every language L_{mf} to which A belongs, A is true for the valuation of that language determined by $<f, \mathscr{E}_m>$.

In dealing with the sequence of languages $L_{1f}, L_{2f}, \ldots$, and the valuations determined, first by f, and then by the $<f, \mathscr{E}_m>$, we have not treated '*I believe that*' as meaning *I explicitly believe that*. The truth-conditional valuations have, in effect, construed '*I believe that*' as '*I am committed to believe that*,' for a sentence '*I believe that A*' comes out true just in case A has value $+$. It is quite natural to describe a person as believing what he is committed to accept, and can reach from explicit beliefs by a short chain of elementary inferences. We are idealizing, and considering the designated subject to believe whatever he is committed to accept. However, actually accepting 'I am committed to believe that *A*' comes to nearly the same as accepting 'I explicitly believe that *A*.'

An initial commitment valuation "registers" those non-modal statements the designated subject accepts and those he rejects to begin with. The initial commitment valuation determines all the subject's further commitments to accept or reject statements within a given moment. When the designated subject (who is proceeding rationally) comes to accept or reject a further statement, this must be a statement he is committed to accept or reject, and he must accept or reject it as the result of a deductively correct inference/argument. If A is a statement that the designated subject is committed to accept, but he has not yet done this, then A is not one of his explicit beliefs. He won't immediately be prepared to accept '*I believe that A*,' but he is not entitled to reject it. For the designated subject to reject '*I believe that A*' is to reject a commitment to accept *A*--to do this he must believe A false or believe that accepting A is incoherent.

The designated subject can *decline* to accept a statement A that is not an explicit belief, but he cannot use '~ *I believe that A*' or reject '*I believe that A*' to signal that A isn't such a belief without depriving the operator of inferential significance.

The commitment valuations have been designed from the standpoint of an "entirely logical" subject. Once he has acquired his initial beliefs and disbeliefs, by whatever means, he acquires all further beliefs (in the same moment) exclusively on the basis of deductively correct inferences/arguments. He doesn't reject any '*I believe that A*' statements until he shows such statements to produce incoherence when added to his explicit beliefs and disbeliefs. The limitation to entirely logical subjects may be excessive. A real person might be able to simply recognize, for some statement A, that he isn't committed to accept A, even though '*I believe that A*' wouldn't be incoherent with his explicit beliefs and disbeliefs. We could change our account of commitment valuations to accommodate such persons, but this would make the account more complicated. For the present, we consider only entirely logical subjects.

In developing the account of commitment valuations, we have considered an interpreting function f and a commitment valuation $\mathscr{E}$ that ultimately rests on f. Doing this ensures that $\mathscr{E}$ is coherent. But f may not reflect the actual state of affairs. We can consider an interpreting function

f^* and a coherent $\mathscr{E}$ not based on f^*. Then in the valuation determined by $<f^*, \mathscr{E}>$, sentences without modal operators get their values from f^*. '*I believe that A*' statements have their values determined by $\mathscr{E}$, and the remaining sentences have truth-table values. Given such a valuation, a statement A might have value F, but nonetheless be believed by the designated subject, so that '*I believe that A*' has value T.

However, in evaluating arguments made by the designated subject, we will presume that his current beliefs are true. We wish to determine which are the arguments that will preserve truth, presuming that his beliefs so far are true. This is important with respect to arguments in which the designated subject supposes true or supposes false certain statements. To evaluate an argument composed entirely of acceptances/assertions and denials, it is enough to consider commitment values. But to suppose A is not to suppose that A is believed, or even that there is a commitment to believe A. If A contains no modal operators, supposing A is supposing A true. To suppose '*I believe that A*' is to suppose that the designated subject is (currently) committed to accept A; to suppose the statement false is to suppose that he isn't committed to accept it.

Implication is the special case of entailment which is identified in terms of the logical forms of sentences in artificial logical languages. We will define implication only for positive situations where statements are accepted/asserted or supposed true. It is easy enough to extend the definitions to cover denials and supposings false. Let $A_1,..., A_n, B$ be plain sentences of L_f. The sentences $A_1,..., A_n$ *basically imply* B iff for every final commitment valuation for which each of $A_1,..., A_n$ has value +, B also has value +.

Suppositional implication is a basically enthymematic notion. Let $\mathscr{E}_0$ be an initial commitment valuation, and $\mathscr{E}$ be the final commitment valuation determined by $\mathscr{E}_0$. Then plain sentences $A_1,..., A_n$ *suppositionally imply B with respect to* $\mathscr{E}_0$ iff for every interpreting function f^* and final commitment valuation $\mathscr{F}$ which extends $\mathscr{E}$, if each of $A_1,..., A_n$ has T for $<f^*, \mathscr{F}>$, then so does B.

Our account of commitment semantics is formulated with respect to the designated subject. Let us consider what it is for an argument made by the designated subject to be *deductively correct*. A simple argument from premisses $?A_1,... ?A_n$ to conclusion $?B$, where each question mark is either '$\vdash$' or '$\llcorner$,' is *deductively correct* iff the premiss acts with their forces commit the designated subject to perform the conclusion act with its force. If all of the question marks are '$\vdash$,' then the argument is deductively correct iff $A_1,..., A_n$ basically imply B. But some or all of the question marks might be '$\llcorner$,' in which case the commitment relations will be different. For an argument from initial assertions $A_1,..., A_m$ and hypotheses (initial suppositions) $B_1,..., B_n$ to conclusion $\llcorner C$, we consider initial commitment valuations $\mathscr{E}_0$ and the final valuation $\mathscr{E}$ determined by $\mathscr{E}_0$ such that each of $A_1,..., A_m$ has value + for $\mathscr{E}$. The sentences $B_1,..., B_n$ must suppositionally imply C with respect to $\mathscr{E}_0$ for each such $\mathscr{E}_0$. A complex argument is *deductively correct* iff its component arguments are deductively correct, and the uncancelled hypotheses and initial assertions of the overall argument commit the subject to perform the conclusion act with its force.

For the designated subject, both of these arguments are deductively correct:

⊢*A*	⊢*I believe that A*
--------------------	--------------------
⊢*I believe that A*	⊢*A*

For the left argument, the premiss act by itself makes the conclusion true. But if the premiss is true, so is the conclusion. The second argument is more interesting. So long as the designated subject accepts the premiss, he is committed to accept the conclusion. But this argument could have a true premiss and a false conclusion. However, each person is committed, for each of his explicit beliefs, to accept the statement of his belief and hold that this belief is true. If the designated subject accepts *A*, and is right about the truth of *A*, then the second argument will take him from a true premiss to a true conclusion. The following arguments are not deductively correct for the designated subject:

⎵*A*	⎵*I believe that A*
--------------------	--------------------
⎵*I believe that A*	⎵*A*

Supposing *A* to be true does not commit the subject to supposing himself committed to accept *A*. And even though the designated subject will think each of his present beliefs to be true, he doesn't think of himself as infallible. So if he supposes himself to believe *A* (to be committed to accept *A*), this doesn't commit him to supposing that *A* is true.

The arguments which are deductively correct for the designated subject are also deductively correct for each of us, if we put ourselves in the position of the designated subject. The deductively correct arguments are deductively correct for whoever is the *I* of the argument. However, an argument which has the same "propositional content" as an argument which is deductively correct for the designated subject can be deductively incorrect for another person. Suppose that Smith is the designated subject. Smith's argument:

⊢*A*

⊢ *I believe that A*

has the same propositional content as this argument made by someone else, or even by Smith himself:

⊢*A*

⊢*Smith believes that A*

But if someone other than Smith makes the argument, it is deductively incorrect for the someone else. And so is:

$$\frac{\vdash \textit{Smith believes that } A}{\vdash A}$$

The argument which is deductively correct for Smith and no one else has as its counterpart the facts that these acts:

$$\vdash A \qquad\qquad \vdash\sim\!\textit{Smith believes that } A$$

are incoherent for Smith to perform, but not for someone else.

Some true statements cannot coherently be accepted by some people. Smith knows (or he should) that no statement of the form '*[A & ~I believe that A]*' can ever be justifiably accepted, even though some such statements may be true. Smith will, however, always be justified in rejecting a statement '*I believe that [A & ~I believe that A].*' With respect to the surprise execution, the moral of our discussion is just that some arguments can be deductively correct for one person, and incorrect for a different person. Certain beliefs can be incoherent for one person, but pose no problem for someone else. And true statements cannot always coherently be believed.

7. BACK TO THE EXECUTION The general strategy for applying the apparatus I have developed to the surprise execution may already be clear. The judge's sentence causes the prisoner to have incoherent beliefs. The prisoner reasons correctly, but he does not, and cannot, arrive at a reliable conclusion. The beliefs which are incoherent for the prisoner are not incoherent for the judge, or for us. However, this application is not entirely straightforward. It will be instructive to see what is involved in leading the prisoner's argument to grief.

For our analysis of arguments, the prisoner will be the designated subject. The prisoner is Smith. We will illustrate and analyze Smith's reasoning, which is carried out subsequent to the judge's verdict but prior to the seven days on one of which the execution is to take place. Although I will sometimes talk about what Smith knows or doesn't know, I will understand this knowledge to be justified belief. Smith's justified beliefs include the judge's verdict, certain other principles which we shall explore, and statements/beliefs arrived at by deductively correct reasoning.

Instead of developing a more complex language with names, predicates, quantifiers, etc., I will simply identify certain sentences and specify what they mean (what statements they represent).

The sentences D_i for $1 \leq i \leq 7$, say that Smith is executed at noon on the *ith* day of the week in question.
The sentences A_i for $1 \leq i \leq 7$, say that Smith is alive at 1 pm on the *ith* day.
The modal operators $\Box_i$ for $1 \leq i \leq 7$, are used in such a way that

$$\Box_i P$$

means that Smith knows (has known, will know) *P* before noon on the *ith* day.

The modal operators $\Box_i$ are simpler than the '*I believe that*' operator considered above. For our boxes are used to represent "eternal" statements, and are not indexical to a present time or moment.

Smith accepts the judge's verdict, and is justified in doing so. The verdict beliefs are the following:

$V1 \ \vdash [D_1 \vee \ldots \vee D_7]$
$Vm, n \ \vdash \sim [D_m \ \& \ D_n]$ for $1 \leq m, n \leq 7$, and $m \neq n$
$V2 \ \vdash [D_n \supset \sim \Box_m D_n]$ for $1 \leq m \leq n \leq 7$

In addition to logical principles based on the connectives, the following inference principles are correct (for Smith) for the knows operator:

$\Box_m$ *Introduction*

$$\frac{\vdash P}{\vdash \Box_m P}$$

The premiss assertion occurs before noon on the *mth* day

$\Box_m$ *Elimination*

$$\frac{\vdash \Box_m P}{\vdash P}$$

(T)

$$\frac{\vdash \Box_m P \qquad \vdash \Box_m [P \supset Q]}{\vdash \Box_m P}$$

This argument occurs before noon on day *m*

A presupposition of the principle $\Box_m$ *Introduction* is that Smith is always in a position to know what day and time it is, and that in the present circumstances Smith makes nothing but justified assertions/acceptances. These presuppositions give the principle an empirical character, but the principle with its presuppositions can fairly be regarded as a defining principle of our problem situation. The principle $\Box_m$ *Elimination* presumes that Smith will not be justified in giving up a belief that he was justified in holding. This principle also characterizes our problem situation. None of these principles is correct if the assertion sign is replaced by the supposition sign. Supposing that *P* is true, even before noon on day *m*, does not commit Smith to supposing that *P* is known before noon on day *m*. And supposing that *P* is justifiably believed before noon on day *m* does not commit Smith to supposing that *P* is true. If Smith actually believes *P*, justifiably, then he is committed to accept *P*, but Smith can entertain the possibility of being mistaken. He needn't hold himself to be infallible. Finally, principle *(T)* doesn't work for suppositions, because with assertion the argument *produces* the belief in question, while supposing the truth of the premisses doesn't give reason for thinking the conclusion must also be true.

The following statements will have an axiomatic status for Smith:

A1 $\vdash [\sim [D_1 \text{ v} \ldots \text{ v } D_m] \supset A_m]$ The subscripts of the disjunction are in ascending order,
so that *m* is the maximum.

A2 $\vdash [A_m \supset \sim [D_1 \text{ v} \ldots \text{ v } D_m]]$ The condition is the same.

A2 can fairly be held to state a law of nature, while *A1*, even though it isn't necessary, constitutes a defining principle for the problem situation. We are interested in the puzzle that arises if *A1* is accepted.

Finally, in the situation we envisage, Smith is justified in accepting these empirical principles:

E1 $\vdash [A_m \supset \Box_{m+1} A_m]$ for $1 \leq m \leq 6$

E2 $\vdash [[\Box_{m+1} A_m \ \& \ \Box_{m+1} [A_m \supset D_{m+1}]] \supset \Box_{m+1} D_{m+1}]$

E1 says that if Smith is alive at 1 pm on day *m*, he will know this before noon the next day. He *could* be alive without giving the matter any thought. But it is certainly the case that in his situation, he will be aware of his condition and of the time that he is in this condition. Indeed, *E1* can be regarded as a decision Smith makes to pay attention to whether he is alive at 1 pm on any given day. And *E2* can be regarded as Smith's decision to notice whether he is alive at 1 pm on any given day, and if he knows that being alive at 1 pm on that day is a necessary condition for being executed the next day, to actually infer the conclusion that he will be executed.

Given all this, at some time before the week in question, Smith argues to establish the following results, our theorems. The proofs of these theorems present his arguments.

Theorem 1 $\vdash[A_6 \supset D_7]$

Proof:

$$\dfrac{\dfrac{\dfrac{\overset{x}{\llcorner A_6} \qquad \overset{A6}{\vdash[A_6 \supset \sim[D_1 \text{ v } \ldots \text{ v } D_6]]}}{\llcorner \sim[D_1 \text{ v } \ldots \text{ v } D_6]}\,\supset E \qquad \overset{V1}{\vdash[D_1 \text{ v } \ldots \text{ v } D_7]}}{\llcorner D_7}\,\textit{propositional logic}}{\vdash[A_6 \supset D_7]}\,\supset I\textit{, drop `}\llcorner A_6\textit{'}$$

Theorem 2 $\vdash\Box_m[A_6 \supset D_7]$ for $1 \leq m \leq 7$

Proof

$$\dfrac{\overset{\textit{Thm 1}}{\vdash[A_6 \supset D_7]}}{\vdash\Box_m[A_6 \supset D_7]}\,\Box_m I$$

The argument takes place prior to day 1.

Theorem 3 $\vdash[D_7 \supset \Box_7 A_6]$

Proof

$$\dfrac{\dfrac{\dfrac{\dfrac{\overset{x}{\llcorner D_7}}{\llcorner \sim[D_1 \text{ v } \ldots \text{ v } D_6]}\,\textit{prop logic, Vm, n} \qquad \overset{A1}{\vdash[\sim[D_1 \text{ v } \ldots \text{ v } D_6] \supset A_6]}}{\llcorner A_6}\,\supset E \qquad \overset{E1}{\vdash[A_6 \supset \Box_7 A_6]}}{\llcorner \Box_7 A_6}\,\supset E}{\vdash[D_7 \supset \Box_7 A_6]}\,\supset I\textit{, drop `}\llcorner D_7\textit{'}$$

Theorem 4 $\vdash[D_7 \supset \Box_7 D_7]$

Proof

$$\dfrac{\dfrac{\dfrac{\dfrac{\overset{x}{\llcorner D_7} \qquad \overset{\mathit{Thm\ 3}}{\vdash[D_7 \supset \Box_7 A_6]}}{\llcorner \Box_7 A_6}\supset E \qquad \overset{\mathit{Thm\ 2}}{\vdash \Box_7[A_6 \supset D_7]}}{\llcorner[\Box_7 A_6 \,\&\, \Box_7[A_6 \supset D_7]]}\,\&I \qquad \overset{E2}{\vdash[[\Box_7 A_6 \,\&\Box_7[A_6 \supset D_7]] \supset \Box_7 D_7]}}{\llcorner \Box_7 D_7}\supset E}{\vdash[D_7 \supset \Box_7 D_7]}\supset I, \text{ drop `}\llcorner D_7\text{'}$$

Theorem 5 $\vdash\sim D_7$

Proof

$$\dfrac{\dfrac{\overset{x}{\llcorner D_7} \qquad \overset{V2}{\vdash[D_7 \supset \sim\Box_7 D_7]}}{\llcorner\sim\Box_7 D_7}\supset E \qquad \dfrac{\overset{x}{\llcorner D_7} \qquad \overset{\mathit{Thm\ 3}}{\vdash[D_7 \supset \Box_7 D_7]}}{\llcorner \Box_7 D_7}\supset E}{\vdash\sim \mathrm{D}_7}\sim I, \text{ drop `}\llcorner D_7\text{'}$$

Given Theorem 5 and *V1*, it is easy to establish this result:

Theorem 6 $\vdash[D_1 \text{ v} \ldots \text{ v } D_6]$

By essentially similar reasoning, which reasoning is deductively correct, Smith can rule out each day as day for his execution. So he ends up with incoherent, even inconsistent, beliefs. He believes that he will be executed on one of the seven days, and he believes that he will be executed on none of those days.

When a person finds himself committed to incoherent assertions, he knows that some of his beliefs need to be given up. There is a fundamental come-what-may commitment to achieve coherence. But here Smith is in trouble, for he has no principled basis for selecting beliefs to be abandoned. If he gives up the belief that he will be executed on one of the seven days, he can achieve coherence. He can also achieve coherence if he gives up the belief that his execution will come as a surprise. Although Smith is rationally required to achieve coherence, reason will not pick out a particular belief to be eliminated.

It is a rational requirement to achieve coherence in one's beliefs, but Smith is required to abandon some beliefs without adopting the opposite beliefs. After all, each of Smith's original beliefs is true. Smith will not be mistaken if he declines to accept a true statement, but will go wrong if he believes the opposite. Smith is in the unfortunate position of having justified true beliefs that he cannot coherently hold.

The predicament that Smith is in is caused by the judge's sentence and by her making it known to him. This, essentially, is what is said in *Thalos 1997*. Smith and the judge have conflicting goals, and the judge has the upper hand. But Thalos' discussion provides no analysis of Smith's reasoning, and no detailed account of the way that Smith goes wrong. Smith's initial beliefs are true, but it is irrational of him to continue to hold these beliefs, for they are not coherent. However, these same beliefs (beliefs with the same "content") pose no problems for us. We can coherently accept the statements that give Smith so much trouble. The difficulty is that some true statements cannot coherently be accepted by some people. The statements in question are off-limits to Smith but not to the rest of us.

However, we should notice that our understanding of Smith's predicament is not beyond Smith's grasp. With a little instruction in illocutionary logic, he himself can realize that his beliefs are incoherent, and that this incoherence is brought about by the judge's sentence. We can all understand what it is for one of our own beliefs to be false even while we think it true. And we can understand how a statement we don't believe might after all be true. If I agree that *A* could be true, even though I don't believe it, then I can agree that '*[A & ~I believe that A]*' might be true. In a similar fashion, Smith can say to himself, "It would be just my luck for all my beliefs to be true, even though I can't coherently hold them."

This kind of reflection won't lead to a new justified belief, but it would certainly make sense for Smith to expect the worst. In that situation, Smith might want some way of determining on which days he can be executed by surprise. Since he can't be sure that he will be executed at all, his execution on any day should be something of a surprise. But his execution will be more of a surprise on any day except the last, than it will be on the last day. And this is a result that Shapiro's procedures would lead us to expect.

John T Kearns
Department of Philosophy and Center for Cognitive Science
State University of New York at Buffalo

REFERENCES

J. L. Austin. 1965. *How To Do Things with Words*. New York: Oxford.

John T. Kearns. 1997. “Propositional Logic of Supposition and Assertion,” *Notre Dame Journal of Formal Logic* 38, pp. 325-349.

G. E. Moore. 1944. “Moore’s Paradox,” in *G. E. Moore, Selected Writings*, edited by Thomas Baldwin, pp. 207-212. London: Routledge, 1993.

Stuart Shapiro. 1998. “A Procedural Solution to the Unexpected Hanging and Sorites Paradoxes,” *Mind* 107, pp. 751-761.

Mariam Thalos. 1997. “Conflict and Coordination in the Aftermath of Oracular Statements,”*The Philosophical Quarterly* 47, pp. 212-226.