KRISTER SEGERBERG

# THE LOGIC OF DELIBERATE ACTION

## 1. PRE-FORMAL NOTIONS

Two stages are discernible in deliberate action, deliberation and performance. The agent decides on what to do, then does it. Or with a more careful formulation: the agent forms an intention, and then tries to carry it out. Deliberation and performance may be complex or not, may take much or little time, may be conscious or less than conscious. Some examples will illuminate this way of viewing action.

EXAMPLE 1. *I want to close the door. So I get up and close the door.*

This is a typical example of a kind of action that we perform every day; actions of this kind make up a considerable portion of our lives. The deliberation involved is minimal: I wish to close the door, there is no overriding consideration to the contrary, and so an intention to close the door (somehow) forms. Performance is unproblematic – closing doors I know something about, especially this door which is the door of my own study and which I have closed innumerable times before. One may say that I have a *routine* for closing this door, and whenever I am closing this door in the normal way I just did, it is this routine that I am running. The routine is not easy to describe, perhaps not even possible to describe. But if you were here, it would be easy to show you: *this* is how I do it. Normally I do it without thinking, sometimes without being aware of doing it. Compare me with the neighbour's one-and-a-half-year-old who might also be able to close the door but who has not yet developed a routine for doing it: he would go about it laboriously, tackling this task with the freshness of a young and untried mind, as yet a *tabula rasa* as far as closing doors is concerned.

One feature of this example is of particular interest: an intention is formed which the agent is able to execute immediately. Let us call such an intention *operational*. Pre-formal theorem: an intention is operational only if there is a routine for trying to carry it out.

Asserting that the performance of my door closing act is in one sense

*Journal of Philosophical Logic* 11 (1982) 233–254. 0022–3611/82/0112–0233$02.20.

unproblematic is not to deny that it is an extremely complicated thing: a human being is a miracle even from an engineering point of view, something you would quickly discover if you would try to build a robot in my effigy, one that would look like me, walk like me, and designed to get up and close the door like me if called upon to do so. This complexity is responsible for the fact that, strictly speaking, the result of my door closing routine varies a little each time it is given a run. That is to say, the sum total of my bodily movements is different each time I close the door, even though the essential thing, that the door gets closed, is always the same. To be certain, these differences are usually of so little importance that it is difficult to see them if you are not a philosopher. Under normal conditions there is no reason why anyone would be interested in the precise locus of my body's centre of gravity during the seconds it takes to close the door. But there are cases when minute differences of this sort matter, differences not due to any indeterminacy in the operational intention but to the variance or imperfection in the routine called up to execute it. Here is a particularly obvious example:

EXAMPLE 2. *I am about to throw a dart. The intention is "to hit the Bull's Eye". I am not very good at this sort of thing, but I have done it often enough, and this is how I do it: I aim carefully, inhale deeply, exhale half my breath, keep the rest, hope for the best, 1, 2, 3, and off it goes!*

I never know where the dart will land, have no sense of how the throw "feels" (which an expert might have). In this respect the dart example differs from the preceding one: I am always confident that the door will be closed at the end of my door closing routine (if there is not some anomaly afoot). Perhaps one may say that I am an expert at closing doors but an amateur at throwing darts.

The observation I am making is a familiar one, but it is of fundamental importance and so will bear repeating once more. In each of the two examples it may be said that I am doing the same thing each time I perform the action, but also that I am doing a different thing each time. If this is difficult to see in the door closing example, it is all the clearer in the dart throwing one. In order to see that I am *doing a different thing* each time I throw the dart, just look at the score: now (usually) it is a bad miss, now (sometimes) it is a near miss, now (once in a long while) it is the Bull's Eye. But it is also true that I am *doing the same thing* each time I throw the dart,

viz., throwing the dart with the intention of hitting the Bull's Eye. Obviously there must be an important distinction waiting to be made here. A statistician would say that it is the same *experiment* that is being repeated but that the *outcomes* differ. Using the vocabulary introduced above, we may add to this by saying that the reason that it is the same experiment is that it is the same routine that is being used.

Both Example 1 and Example 2 involve no or little deliberation. Here is a more complicated example involving considerable deliberation:

EXAMPLE 3. *I wish to give X a birthday present. What will it be: flowers, bottle of wine, a book? Knowing X, I decide: a book. But what book? For various reasons, who knows which, I decide: a book by Fritiof Nilsson the Pirate. Surveying the contents of my bookstore I see a copy of* The book dealer who gave up bathing, *and I decide: that book. So I buy the copy and have it sent to X.*

It is instructive to follow the stages of deliberation in the example. A series of, as it happens, strictly monotonically more specific intentions is formed until one is reached which can be realized immediately – an intention that is operational. Notice that in the chain of intentions formed in the example none except the last one is operational: "to give X a birthday present", "to give X a book", "to give X a book by Fritiof Nilsson the Pirate" were all intentions of mine, but none of them could be carried out immediately. By contrast, the last intention, "to give X *The book dealer who gave up bathing*" could be realized at once and in a simple way. In a generous sense of routine, there was a routine which I could associate with this intention. There was no compulsion to run this routine, but as a matter of fact I did run it.

A final observation. Looking back on what I did in this example, it would be true to say that I have given X a birthday present, that I have given X a book, that I have given X a book by the Pirate, etc. But would it be true to assert that I have given X a book or a record (in the sense in which to give something is to give it intentionally)? In one sense such an assertion would seem to be true, for by classical logic, if I have given X a book, then either I have given X a book or I have given X a record. But there is also a sense in which the assertion would seem to be false, for by assumption "to give X a book or a record" was never an intention of mine.

The sketchiness of the preceding discussion will no doubt leave philosophers of action dissatisfied. However, as a pre-theoretical background to what follows it will suffice. The main purpose of the paper is to provide a completeness result for the intentional logic first defined in [2]. Actually, the logic presented here is slightly more general than the one defined there, but the completeness claim made there is an easy consequence of the work done here.

## 2. FORMAL NOTIONS

We begin by laying down a formal semantics meant to model deliberate action as conceived above. An *outcome space* $U$ is a non-empty set, the elements of which are called *outcomes* and the subsets of which are called *events* (note that the concepts of outcome and event are both relative to a given space). Two events stand out, *the impossible event* $\emptyset$ and *the certain event* $U$. An *action* in $U$ is a pair $\langle S, x\rangle$ such that $S \subseteq \mathfrak{P}\, U$ and $x \in U$; that is, $S$ is a set of events and $x$ is an outcome. By an *intentional action structure (based on U)* we mean any non-empty family of actions in $U$.

The intuitive considerations in Section 1 should provide some motivation for our choice of formalism. If the defintion of action appears artificial, let the following remarks be added. We think of an intention as an intention to bring about an event, and thus it is convenient, in our modelling, to identify an intention with the corresponding event. Furthermore, the agent (who is tacitly understood and fixed throughout the formal development) is thought to have a definite set $S$ of intentions in mind which determines "what he does"; Example 3 shows that this set may have more than one element. But as shown by Example 2, the agent does not always completely control the world around him, hence we need to know the outcome $x$ before we know exactly "what he did". Another way to think about this is the following. We wish to find a formal representation of action. If you are told that the construct $\langle S, x\rangle$ models a certain particular action, then you know all there is to know about this action *per se* (within the present context): knowing $S$ you know what the agent intended by his action, and knowing $x$ you know what actually happened. Admittedly, this is a crude theory, and, as will be seen in Section 5, narrowly limited. Yet for the logic of action it seems to the author to promise something of a new beginning.

To fit this semantics with an object language we proceed as follows. First

TABLE I

| Category | Symbols | Appellation |
|---|---|---|
| $\mathbf{t}$ | $\pi_0, \pi_1, \ldots, \pi_n, \ldots$ | event letters |
| $\mathbf{t}^{\mathbf{t}}$ | $-$ | event complement |
| $\mathbf{t}^{\mathbf{t},\mathbf{t}}$ | $\frown$ | event intersection |
| | $\smile$ | event union |
| $\mathbf{f}$ | $\mathbf{P}_0, \mathbf{P}_1, \ldots, \mathbf{P}_n, \ldots$ | propositional letters |
| $\mathbf{f}^{\mathbf{f}}$ | $\neg$ | negation |
| | $\boxdot$ | necessity |
| $\mathbf{f}^{\mathbf{f},\mathbf{f}}$ | $\wedge$ | conjunction |
| | $\vee$ | disjunction |
| | $\rightarrow$ | material implication |
| | $\leftrightarrow$ | material equivalence |
| $\mathbf{f}^{\mathbf{t}}$ | Int | the intention operator |
| | Real | the realization operator |
| $\mathbf{f}^{\mathbf{t},\mathbf{t}}$ | $=$ | event identity |

we introduce a categorical notation of the kind used by Ajdukiewicz and Montague and also used in a similar context by the author in [1]. Let $\mathbf{t}$ and $\mathbf{f}$ be the two basic categories ($\mathbf{t}$ for "term" and $\mathbf{f}$ for "formula"). Then we shall have primitive symbols as listed in Table I. We assume of course that the intersection of the sets of primitive symbols of any two categories, basic or derived, is empty. Notice that the table completely determines our syntax. In particular it defines the class of expressions of category $\mathbf{f}$ (*formulas*) and the class of expressions of category $\mathbf{t}$ (*terms*). The simplest terms are the event letters, while any complex term is of the form $\bar{\alpha}$, $\alpha \frown \beta$, or $\alpha \smile \beta$. The simplest formulas are the propositional letters, but there are also others that are simple in the sense that they do not contain any formula as a proper sub-expression, viz., formulas of type Int $\alpha$, Real $\alpha$, and $\alpha = \beta$. Every formula that is not simple in this sense is of the form $\neg\mathbf{A}$, $\boxdot\mathbf{A}$, $\mathbf{A} \wedge \mathbf{B}$, $\mathbf{A} \vee \mathbf{B}$, $\mathbf{A} \rightarrow \mathbf{B}$, or $\mathbf{A} \leftrightarrow \mathbf{B}$. Meta-logical conventions: we will use $\mathbf{P}$ for propositional letters; $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ for general formulas; $\pi$ for event letters; $\alpha, \beta, \gamma$ for general terms. Formulas of type $\alpha = \beta$ we sometimes call *equations*. We introduce as abbreviatory devices three operators: $\neq, \leq,$ $\Diamond$ of category $\mathbf{f}^{\mathbf{t},\mathbf{t}}$, $\mathbf{f}^{\mathbf{t},\mathbf{t}}$, respectively $\mathbf{f}^{\mathbf{f}}$:

"$\alpha \neq \beta$" for $\neg(\alpha = \beta)$,

"$\alpha \leq \beta$" for $\alpha \frown \beta = \alpha$

"$\Diamond\mathbf{A}$" for $\neg\boxdot\neg\mathbf{A}$.

Three of the primitive operators need special comment: $\boxdot$, Int and Real. As suggested above, $\boxdot$ is seen as a necessity operator in the sense of modern modal logic, but it is not logical necessity that is meant. The reading of $\boxdot\mathbf{A}$ favoured in [2] was "it is part of the situation that **A**". Both Int and Real operate on terms, and Int $\alpha$ and Real $\alpha$ may be read "$\alpha$ is intended by the agent", respectively, "$\alpha$ is realized". These suggested readings are of course unofficial: the meaning of the operators is given by the truth conditions that will now be given.

Let $\mathfrak{S}$ be an intentional action structure based on some set $U$. Then we say that $\mathfrak{M} = \langle U, \mathfrak{S}, V\rangle$ is a *model* (*in* $\mathfrak{S}$) if $V$ is a valuation; that is, if $V$ is a function assigning to each event letter a subset of $U$ and to each propositional letter a truth-value, either T (truth) or F (falsity).

The first thing to notice is that $\mathfrak{M}$ in effect assigns to each term $\alpha$ an event $\|\alpha\|^{\mathfrak{M}}$ (we will drop the superscript wherever possible):

$$\|\pi\| = V(\pi),$$

$$\|\mathbf{1}\| = U,$$

$$\|\mathbf{0}\| = \emptyset,$$

$$\|\alpha \frown \beta\| = \|\alpha\| \cap \|\beta\|,$$

$$\|\alpha \smile \beta\| = \|\alpha\| \cup \|\beta\|,$$

$$\|\bar{\alpha}\| = U - \|\alpha\|.$$

The next thing to notice is that $\mathfrak{M}$ in effect assigns to each formula not involving Int or Real or $\boxdot$ a truth-value. However, we want a truth-value for every formula, and in order to achieve this we have to settle for one that is relativized to an action $\langle S, x\rangle$. In accordance with custom we will write $\langle S, x\rangle \models^{\mathfrak{M}} \mathbf{A}$, if **A** is assigned T, and $\langle S, x\rangle \not\models^{\mathfrak{M}} \mathbf{A}$, if **A** is assigned F. As before, we drop the superscripts wherever possible:

$$\langle S, x\rangle \models \mathbf{P} \quad \text{iff } V(\mathbf{P}) = \mathrm{T},$$

$$\langle S, x\rangle \models \alpha = \beta \quad \text{iff } \|\alpha\| = \|\beta\|,$$

$$\langle S, x\rangle \models \mathsf{Int}\,\alpha \quad \text{iff } \|\alpha\| \in S,$$

$$\langle S, x\rangle \models \mathsf{Real}\,\alpha \quad \text{iff } x \in \|\alpha\|,$$

$$\langle S, x\rangle \models \neg\mathbf{A} \quad \text{iff } \langle S, x\rangle \not\models \mathbf{A},$$

$\langle S, x\rangle \models \mathbf{A} \wedge \mathbf{B}$ iff $\langle S, x\rangle \models \mathbf{A}$ and $\langle S, x\rangle \models \mathbf{B}$,

etc.

$\langle S, x\rangle \models \boxdot \mathbf{A}$ iff, for all $\langle T, y\rangle \in \mathfrak{S}$, $\langle T, y\rangle \models \mathbf{A}$.

Let us say that **A** is *true in* $\mathfrak{M}$ if, for all $\langle S, x\rangle \in \mathfrak{S}$, $\langle S, x\rangle \models^{\mathfrak{M}} \mathbf{A}$; *valid in* $\mathfrak{S}$ if true in all models in $\mathfrak{S}$; *valid* if valid in all intentional action structures. We say that a model $\mathfrak{M} = \langle U, \mathfrak{S}, V\rangle$ is a *model for* a set $\Sigma$ of formulas, or that $\Sigma$ *has* $\mathfrak{M}$ *as a model*, if there is some $\langle S, x\rangle \in \mathfrak{S}$ such that, for all $\mathbf{A} \in \Sigma$, $\langle S, x\rangle \models^{\mathfrak{M}} \mathbf{A}$.

We turn now to the problem of axiomatizing the set of valid formulas. To this end we lay down the following somewhat redundant axiom system:

AXIOMS

(BA) All valid equations of Boolean algebra.

(TF) All instances of truth-functional tautologies.

(S5) All instances of formulas valid in Lewis' system S5.

(N1) $\mathbf{A} \rightarrow \boxdot \mathbf{A}$, if **A** contains no occurrence of Int or Real.

(E1) $\alpha = \alpha$.

(E2) $\alpha = \beta \rightarrow (\mathbf{A} \leftrightarrow \mathbf{A}')$, if **A** and $\mathbf{A}'$ are similar formulas differing only in one place where **A** has an occurrence of $\alpha$ and $\mathbf{A}'$ an occurrence of $\beta$.

(R0) $\neg$ Real 0.

(R1) Real 1.

(R2) Real $(\alpha \frown \beta) \leftrightarrow$ (Real $\alpha \wedge$ Real $\beta$).

(R3) Real $(\alpha \smile \beta) \leftrightarrow$ (Real $\alpha \vee$ Real $\beta$).

INFERENCE RULES

(Modus Ponens) $\mathbf{A}, \mathbf{A} \rightarrow \mathbf{B}/\mathbf{B}$.

(Necessitation) $\mathbf{A}/\boxdot \mathbf{A}$

Here (N1), (E1), (E2), (R2), and (R3) – but not (R0) and (R1) – are axiom

schemata, each covering infinitely many formulas. For example, according to (N1) every formula $\mathbf{A} \to \Box\mathbf{A}$ is an axiom if $\mathbf{A}$ is a formula not containing Int or Real, and according to (E1) every formula of type $\alpha = \alpha$ is an axiom.

By M.I.L., the *Minimal Intentional Logic*, we mean the smallest set of formulas which contains all of the above axioms and which is also closed under the two inference rules. We write $\mathbf{A}_0, \ldots, \mathbf{A}_{n-1} \vdash \mathbf{B}$ if the formula $(\mathbf{A}_0 \wedge \ldots \wedge \mathbf{A}_{n-1}) \to \mathbf{B}$ belongs to M.I.L., and $\Sigma \vdash \mathbf{B}$ if $\Sigma$ is a set of formulas and there is some $n \geq 0$ such that $\mathbf{A}_0, \ldots, \mathbf{A}_{n-1} \vdash \mathbf{B}$, for some $\mathbf{A}_0, \ldots, \mathbf{A}_{n-1} \in \Sigma$. Furthermore, $\Sigma$ is said to be *consistent* if there is at least one formula $\mathbf{B}$ such that not $\Sigma \vdash \mathbf{B}$.

Axiom schema (N1) is very strong. Without the proviso the modal aspect of our logic would be trivial. But as the presence of the proviso suggests, M.I.L. is not closed under substitution of formulas for propositional letters; even though all formulas of type $\mathbf{P} \to \Box\mathbf{P}$ are theorems of M.I.L., formulas of type $\mathsf{Int}\,\alpha \to \Box\mathsf{Int}\,\alpha$ or $\mathsf{Real}\,\alpha \to \Box\mathsf{Real}\,\alpha$ usually are not. The latter claim is a corollary of the completeness theorem which is stated below at the end of this section.

The strength of axiom schema (E2) is also worth noting. Among its consequences are the following observations which we list for future reference:

(EM) $\quad \alpha = \beta \vdash \Box(\alpha = \beta)$,

(EI) $\quad \alpha = \beta \vdash \mathsf{Int}\,\alpha \leftrightarrow \mathsf{Int}\,\beta$,

(ER) $\quad \alpha = \beta \vdash \mathsf{Real}\,\alpha \leftrightarrow \mathsf{Real}\,\beta$.

A logic of this kind was first discussed in [2], where it was also argued that such a logic can afford a useful analysis of some notions of agency and ability. Thus to express that *the agent intentionally does* $\alpha$, there are the two prime candidates

$$\mathsf{Int}\,\alpha \wedge \Box(\mathsf{Int}\,\alpha \to \mathsf{Real}\,\alpha),$$

$$\mathsf{Int}\,\alpha \wedge \mathsf{Real}\,\alpha,$$

and to express that *the agent can do* $\alpha$ there are the corresponding two prime candidates

$$\Diamond\mathsf{Int}\,\alpha \wedge \Box(\mathsf{Int}\,\alpha \to \mathsf{Real}\,\alpha),$$

$$\Diamond(\mathsf{Int}\,\alpha \wedge \mathsf{Real}\,\alpha).$$

The differences between the operators Int and Real should be noted. Axioms (R0) and (R1) are syntactic reflexions of the fact that in an intentional action structure, Real 0 is false and Real 1 true. On the other hand, Int 0 and Int 1 can be true or false in models, and accordingly there are no counterparts of (R0) or (R1) among our axioms. This agrees with pre-formal intuitions: an agent may well attempt the impossible or the certain, even if he usually does not. At the same time it must be noted that our modelling is crude and, for one thing, does not do justice to epistemic considerations. Thus an agent attempting to do the impossible or the certain is usually not aware that this is what he is doing, but this nuance cannot be rendered in our system. We return to this topic at the end of the paper.

In the same vein it should be noted that we always have

$$\alpha \leq \beta, \mathsf{Real}\,\alpha \vdash \mathsf{Real}\,\beta,$$

but *not* in general

$$\alpha \leq \beta, \mathsf{Int}\,\alpha \vdash \mathsf{Int}\,\beta.$$

Again this is in agreement with pre-formal intuitions.

The scene is now set for the completeness theorem:

THEOREM. *Every consistent set of formulas has a model.*

As remarked before, the main purpose of this paper is to provide a proof of this theorem.

## 3. THE COMPLETENESS PROOF

The gist of the proof is a construction which we will now describe. It is important to note that it is relative to a maximal consistent set. That is to say, given any maximal consistent set $\Sigma$, the construction yields a certain model $\mathfrak{M}_\Sigma$ with a very special property described in Lemma 3 below. The construction is yet another and somewhat interesting variation of an age-old theme, the Lindenbaum/Stone/Henkin/Tarski/Jónsson Evergreen.

Let $\Sigma$ be a given fixed maximal consistent set of formulas. The following defines a binary relation in the set $\Theta$ of terms:

$$\alpha \sim \beta \,(mod\ \Sigma) \quad \text{iff } \alpha = \beta \in \Sigma.$$

This relation is called the *equivalence relation induced by* $\Sigma$. The

terminology is not arbitrary, for with the aid of axioms from (TF), (E1), and (E2) it is easy to see that $\sim (mod\ \Sigma)$ is reflexive, symmetric, and transitive. Thus we may define

$$\alpha/\Sigma =_{df} \{\beta\colon \alpha \sim \beta\ (mod\ \Sigma)\}$$

and know that $\alpha/\Sigma = \beta/\Sigma$ if and only if $\alpha = \beta \in \Sigma$. In the same way it can be shown that

$$\alpha \sim \alpha'\ (mod\ \Sigma) \quad \text{and} \quad \beta \sim \beta'\ (mod\ \Sigma)$$
$$\text{only if } \alpha \frown \beta \sim \alpha' \frown \beta'\ (mod\ \Sigma),$$

$$\alpha \sim \alpha'\ (mod\ \Sigma) \quad \text{and} \quad \beta \sim \beta'\ (mod\ \Sigma)$$
$$\text{only if } \alpha \smile \beta \sim \alpha' \smile \beta'\ (mod\ \Sigma),$$

$$\alpha \sim \beta\ (mod\ \Sigma) \quad \text{only if } \bar{\alpha} \sim \bar{\beta}\ (mod\ \Sigma),$$

$$\alpha \sim \beta\ (mod\ \Sigma) \quad \text{only if } \mathsf{Int}\,\alpha \in \Sigma \text{ iff } \mathsf{Int}\,\beta \in \Sigma,$$

$$\alpha \sim \beta\ (mod\ \Sigma) \quad \text{only if } \mathsf{Real}\,\alpha \in \Sigma \text{ iff } \mathsf{Real}\,\beta \in \Sigma.$$

We define the *Lindenbaum algebra* (*for* $\Sigma$) as the structure $\mathfrak{A}_\Sigma = \langle \Theta/\Sigma, \frown, \smile, -, \mathbb{0}_\Sigma, \mathbb{1}_\Sigma\rangle$, where

$$\Theta/\Sigma = \{\alpha/\Sigma\colon \alpha \in \Theta\},$$

$$\alpha/\Sigma \frown \beta/\Sigma = \alpha \frown \beta/\Sigma,$$

$$\alpha/\Sigma \smile \beta/\Sigma = \alpha \smile \beta/\Sigma,$$

$$-(\alpha/\Sigma) = \bar{\alpha}/\Sigma,$$

$$\mathbb{0}_\Sigma = 0/\Sigma,$$

$$\mathbb{1}_\Sigma = 1/\Sigma.$$

Preceding remarks guarantee that the definition is meaningful. Let the set $U_\Sigma$ of all ultrafilters in $\mathfrak{A}_\Sigma$ be called the *canonical outcome space* and introduce the notation

$$|\alpha|_\Sigma = \{x \in U_\Sigma\colon \alpha/\Sigma \in x\}.$$

This notation is correct, and $|\alpha|_\Sigma = |\beta|_\Sigma$ if and only if $\alpha \sim \beta\ (mod\ \Sigma)$. We define the *canonical intentional action structure on* $U_\Sigma$ as the set $\mathfrak{S}_\Sigma$ of all pairs $\langle S, x\rangle$ where $S \subseteq \mathfrak{P}U_\Sigma$ and $x \in U_\Sigma$ that satisfy the following condition:

for all natural numbers $m, n$ and terms $\alpha_0, \ldots, \alpha_{m-1}, \beta_0, \ldots, \beta_{n-1}, \gamma$, if $|\alpha_i|_\Sigma \in S$, for all $i < m$, and $|\beta_i|_\Sigma \notin S$, for all $i < n$, and, furthermore, $x \in |\gamma|_\Sigma$, then

$$\Diamond(\mathsf{Int}\,\alpha_0 \wedge \ldots \wedge \mathsf{Int}\,\alpha_{m-1} \wedge \neg\mathsf{Int}\,\beta_0 \wedge \ldots \wedge \neg\mathsf{Int}\,\beta_{n-1} \wedge \mathsf{Real}\,\gamma) \in \Sigma.$$

The *canonical valuation* $V_\Sigma$ is the function defined on the set of all event letters and propositional letters such that

$$V_\Sigma(\pi) = |\pi|_\Sigma,$$
$$V_\Sigma(\mathbf{P}) = \mathrm{T} \quad \text{iff } \mathbf{P} \in \Sigma.$$

The *canonical model*, finally, is the triple $\mathfrak{M}_\Sigma = \langle U_\Sigma, \mathfrak{S}_\Sigma, V_\Sigma \rangle$.

Before proceeding we still need a few technical definitions. We will say that two maximal consistent sets $\Gamma$ and $\Gamma'$ of formulas are *statically equi-equivalent* if they agree on modalized formulas:

$$\Gamma \simeq \Gamma' \quad \text{iff } \forall \mathbf{A}(\boxdot\mathbf{A} \in \Gamma \text{ iff } \boxdot\mathbf{A} \in \Gamma').$$

It is clear that $\simeq$ is an equivalence relation. Note that it follows from the definition of $\simeq$ together with axioms from (S5), (N1) and (EM) that statically equivalent maximal consistent sets agree also on propositional letters and equations; that is, whenever $\Gamma \simeq \Gamma'$, then $\mathbf{P} \in \Gamma$ iff $\mathbf{P} \in \Gamma'$, and $\alpha = \beta \in \Gamma$ iff $\alpha = \beta \in \Gamma'$. In fact, statically equivalent sets disagree only on Boolean combinations that involve formulas of type $\mathsf{Int}\,\alpha$ and $\mathsf{Real}\,\alpha$.

Let us say that a maximal consistent set $\Gamma$ *fits* a pair $\langle S, x \rangle \in \mathfrak{S}_\Sigma$ (*over* $\Sigma$) if the following three conditions are satisfied:

(i) $\Gamma \simeq \Sigma$,

(ii) $|\alpha|_\Sigma \in S$ iff $\mathsf{Int}\,\alpha \in \Gamma$, for all $\alpha$,

(iii) $x \in |\gamma|_\Sigma$ iff $\mathsf{Real}\,\gamma \in \Gamma$, for all $\gamma$.

It is important to note that the notion of fit is relative to the fixed set $\Sigma$.

We have now introduced all the conceptual machinery needed for our venture. From a heuristic point of view the reader would be well advised at this point to look ahead at Lemma 4, including its proof and the subsequent remarks. If he does this, he will understand the rationale for our particular definition of canonical model and why the untraditional notion of fit was developed, and he will also realize why Lemma 1 and 2 below are needed.

LEMMA 1. *For any pair* $\langle S, x\rangle \in \mathfrak{S}_\Sigma$ *there is some maximal consistent set* $\Gamma$ *that fits* $\langle S, x\rangle$ *over* $\Sigma$.

*Proof.* Define the set $\Omega = \Omega_0 \cup \Omega_1 \cup \Omega_2 \cup \Omega_3 \cup \Omega_4$, where

$$\Omega_0 = \{\boxdot\mathbf{A}\colon \boxdot\mathbf{A} \in \Sigma\},$$

$$\Omega_1 = \{\neg\boxdot\mathbf{B}\colon \boxdot\mathbf{B} \notin \Sigma\},$$

$$\Omega_2 = \{\mathsf{Int}\ \alpha\colon |\alpha|_\Sigma \in S\},$$

$$\Omega_3 = \{\neg\mathsf{Int}\ \beta\colon |\beta|_\Sigma \notin S\},$$

$$\Omega_4 = \{\mathsf{Real}\ \gamma\colon x \in |\gamma|_\Sigma\}.$$

The definitions of $\Omega_2$, $\Omega_3$ and $\Omega_4$ are correct, thanks to (Necessitation), (S5), (EI), and (ER). We will now show that $\Omega$ is consistent. Suppose it were not: then for some natural numbers $m, n, p, q, r \geqq 0$ there would be formulas $\boxdot\mathbf{A}_0, \ldots, \boxdot\mathbf{A}_{m-1} \in \Omega_0$, $\neg\boxdot\mathbf{B}_0, \ldots, \neg\boxdot\mathbf{B}_{n-1} \in \Omega_1$, $\mathsf{Int}\ \alpha_0, \ldots, \mathsf{Int}\ \alpha_{p-1} \in \Omega_2$, $\neg\mathsf{Int}\ \beta_0, \ldots, \neg\mathsf{Int}\ \beta_{q-1} \in \Omega_3$, $\mathsf{Real}\ \gamma_0, \ldots, \mathsf{Real}\ \gamma_{r-1} \in \Omega_4$ which are jointly inconsistent. In view of axiom (R1) there is no loss of generality if we assume that $r > 0$ and $\gamma_0 = 1$. Let us write $\mathbf{C}$ for the long conjunction

$$\mathsf{Int}\ \alpha_0 \wedge \ldots \wedge \mathsf{Int}\ \alpha_{p-1} \wedge \neg\mathsf{Int}\ \beta_0 \wedge \ldots \wedge \neg\mathsf{Int}\ \beta_{q-1} \wedge \mathsf{Real}\ (\gamma_0 \frown \ldots \frown \gamma_{r-1}).$$

Using, among other things, (R2), we conclude that

$$\boxdot\mathbf{A}_0, \ldots, \boxdot\mathbf{A}_{m-1}, \neg\boxdot\mathbf{B}_0, \ldots, \neg\boxdot\mathbf{B}_{n-1} \vdash \neg\mathbf{C}.$$

By (Necessitation), (S5), etc.,

$$\boxdot\mathbf{A}_0, \ldots, \boxdot\mathbf{A}_{m-1}, \neg\boxdot\mathbf{B}_0, \ldots, \neg\boxdot\mathbf{B}_{n-1} \vdash \boxdot\neg\mathbf{C}.$$

Since $\boxdot\mathbf{A}_0, \ldots, \boxdot\mathbf{A}_{m-1}, \neg\boxdot\mathbf{B}_0, \ldots, \neg\boxdot\mathbf{B}_{n-1} \in \Sigma$, it follows that $\boxdot\neg\mathbf{C} \in \Sigma$. But we have assumed that $\langle S, x\rangle \in \mathfrak{S}_\Sigma$. Therefore, since by construction $\gamma_0 \frown \ldots \frown \gamma_{r-1}$ is a non-empty term, it follows that $\diamondsuit\mathbf{C} \in \Sigma$; a contradiction. Consequently, $\Omega$ is consistent.

As $\Omega$ is consistent, it can be extended, by Lindenbaum's Lemma, to some maximal consistent set $\Gamma$. We claim that $\Gamma$ fits $\langle S, x\rangle$. This claim is easy to prove: the only part that is not entirely straightforward is when it comes to proving that, for any term $\alpha$, if $\mathsf{Real}\ \alpha \in \Gamma$ then $x \in |\alpha|_\Sigma$. Here one

must first establish that $\vdash \neg(\mathsf{Real}\ \alpha \wedge \mathsf{Real}\ \bar{\alpha})$; but this is easily done with the help of (R0) and (R2). ∎

LEMMA 2. *For any maximal consistent set* $\Gamma \simeq \Sigma$ *there is some pair* $\langle S, x\rangle \in \mathfrak{S}_\Sigma$ *which* $\Gamma$ *fits over* $\Sigma$.

*Proof.* Consider the set $T = \bigcap\{|\gamma|_\Sigma\colon \mathsf{Real}\ \gamma \in \Gamma\}$. We wish to show that $T \neq \emptyset$. Suppose, by contradiction, that $T = \emptyset$. It is well known that the topology, the open sets of which are the unions of members of the set $\{|\alpha|_\Sigma\colon \alpha \in \Theta\}$, is compact. Moreover, each set $|\alpha|_\Sigma$ is both open and closed in this topology. Therefore, as $T$ is the intersection of a collection of closed sets, there must be some terms $\gamma_0, \ldots, \gamma_{m-1}$ such that

(1) $\quad \mathsf{Real}\ \gamma_0, \ldots, \mathsf{Real}\ \gamma_{m-1} \in \Gamma,$

(2) $\quad |\gamma_0|_\Sigma \cap \ldots \cap |\gamma_{m-1}|_\Sigma = \emptyset.$

Because of axiom (R1) there is no loss of generality if we assume that $m > 0$. Now, (2) implies that $|\gamma_0 \frown \ldots \frown \gamma_{m-1}|_\Sigma = \emptyset$, and so $(\gamma_0 \frown \ldots \frown \gamma_{m-1} = 0) \in \Sigma$. Hence, since by hypothesis $\Gamma \simeq \Sigma$ and so $\Gamma$ and $\Sigma$ agree on equations,

(3) $\quad (\gamma_0 \frown \ldots \frown \gamma_{m-1} = 0) \in \Gamma.$

From (1) and axiom schema (R2) it follows that

(4) $\quad \mathsf{Real}\ (\gamma_0 \frown \ldots \frown \gamma_{m-1}) \in \Gamma.$

By (3), (4) and (ER), $\mathsf{Real}\ 0 \in \Gamma$, contradicting (R0). Thus $T \neq \emptyset$.

Define $S = \{|\alpha|_\Sigma\colon \mathsf{Int}\ \alpha \in \Gamma\}$ and pick any $x \in T$. (It is in order to do this that we need to know that $T \neq \emptyset$.) We contend that $\langle S, x\rangle \in \mathfrak{S}_\Sigma$. Suppose that there are natural numbers $m, n$ and terms $\alpha_0, \ldots, \alpha_{m-1}, \beta_0, \ldots, \beta_{n-1}$, $\gamma$ such that, for all $i < m$, $|\alpha_i|_\Sigma \in S$, and for all $i < n$, $|\beta_i|_\Sigma \notin S$, and $x \in |\gamma|_\Sigma$. Then

(5) $\quad$ for all $i < m$, $\mathsf{Int}\ \alpha_i \in \Gamma,$

(6) $\quad$ for all $i < n$, $\neg\mathsf{Int}\ \beta_i \in \Gamma,$

(7) $\quad \mathsf{Real}\ \gamma \in \Gamma.$

That (7) holds is not obvious, so let us append the following argument. Suppose that $\mathsf{Real}\ \gamma \notin \Gamma$. Then, by (R1), (R3), and (ER), $\mathsf{Real}\ \bar{\gamma} \in \Gamma$, and so

by the definition of $T$ we have $T \subseteq |\bar{\gamma}|_\Sigma$. But $x$ is an ultrafilter, and *a fortiori* a proper filter, and so $x \in |\gamma|_\Sigma$ implies that $x \notin |\bar{\gamma}|_\Sigma$ (for otherwise we would have both $\gamma/\Sigma \in x$ and $\bar{\gamma}/\Sigma \in x$; and so by the filter property $\gamma \frown \bar{\gamma}/\Sigma \in x$, which is the same as $0/\Sigma \in x$; and that would contradict properness). Consequently $x \notin T$, which is a contradiction. Hence (7).

Let us write

$$\mathbf{C} = \mathsf{Int}\,\alpha_0 \wedge \ldots \wedge \mathsf{Int}\,\alpha_{m-1} \wedge \neg\mathsf{Int}\,\beta_0 \wedge \ldots \wedge \neg\mathsf{Int}\,\beta_{n-1} \wedge \mathsf{Real}\,\gamma.$$

By (5), (6), and (7), $\mathbf{C} \in \Gamma$. Therefore $\Diamond\mathbf{C} \in \Gamma$; and so $\Diamond\mathbf{C} \in \Sigma$, since by hypothesis $\Gamma \simeq \Sigma$. This proves the contention that $\langle S, x\rangle \in \mathfrak{S}_\Sigma$.

Finally we claim that $\Gamma$ fits $\langle S, x\rangle$. Condition (i) in the definition of fit is satisfied by hypothesis, condition (ii) by the definition of $S$. For condition (iii), note that, by the definition of $T$, $\mathsf{Real}\,\gamma \in \Gamma$ implies that $x \in |\gamma|_\Sigma$, while – as above – $\mathsf{Real}\,\gamma \notin \Gamma$ implies $\mathsf{Real}\,\bar{\gamma} \in \Gamma$, and so $x \in |\bar{\gamma}|_\Sigma$ and $x \notin |\gamma|_\Sigma$. ∎

LEMMA 3. *For every term* $\alpha$, $V_\Sigma(\alpha) = |\alpha|_\Sigma$.

*Proof.* This is an elementary result which, however, is needed below. An inductive argument readily yields the lemma once the following pre-lemma has been established:

(i) $\quad |\alpha \frown \beta|_\Sigma = |\alpha|_\Sigma \cap |\beta|_\Sigma,$

(ii) $\quad |\alpha \smile \beta|_\Sigma = |\alpha|_\Sigma \cup |\beta|_\Sigma,$

(iii) $\quad |\bar{\alpha}|_\Sigma = U_\Sigma - |\alpha|_\Sigma.$

This pre-lemma in turn is immediate when it is observed that the elements of $U_\Sigma$ are ultrafilters. For example, the argument for (i) goes as follows: for any $x \in U_\Sigma$, $x \in |\alpha \frown \beta|_\Sigma$ iff $\alpha \frown \beta/\Sigma \in x$ iff $\alpha/\Sigma \frown \beta/\Sigma \in x$ iff (by the filter property) $\alpha/\Sigma \in x$ and $\beta/\Sigma \in x$ iff $x \in |\alpha|_\Sigma$ and $x \in |\beta|_\Sigma$ iff $x \in |\alpha|_\Sigma \cap |\beta|_\Sigma$. ∎

LEMMA 4. *Let* $\mathbf{A}$ *be any formula,* $\Gamma$ *any maximal consistent set of formulas,* $\langle S, x\rangle$ *any member of* $\mathfrak{S}_\Sigma$. *Suppose that* $\Gamma$ *fits* $\langle S, x\rangle$ *over* $\Sigma$. *Then*

$$\langle S, x\rangle \models^{\mathfrak{M}_\Sigma} \mathbf{A} \quad \textit{iff}\ \mathbf{A} \in \Gamma.$$

*Proof.* By induction on $\mathbf{A}$. The basic step consists of four cases according to whether $\mathbf{A}$ is a propositional letter or of the form $\alpha = \beta$, $\mathsf{Int}\,\alpha$, or $\mathsf{Real}\,\alpha$.

All four cases follow from the assumption that $\Gamma$ fits $\Sigma$, the first case immediately, the last three mediately *via* Lemma 3.

The inductive step consists of several Boolean cases, which are immediate, and one modal case. The modal case has two parts, and we treat them separately. Common to them is the induction hypothesis that the lemma holds for $\mathbf{A}$.

First suppose that $\langle S, x\rangle \not\models \boxdot\mathbf{A}$. Then $\langle T, y\rangle \not\models \mathbf{A}$, for some $\langle T, y\rangle \in \mathfrak{S}_\Sigma$. By Lemma 1 there is some maximal consistent set $\Delta$ fitting $\langle T, y\rangle$ over $\Sigma$, and so, by the induction hypothesis, $\mathbf{A} \notin \Delta$. Consequently, by (S5), $\boxdot\mathbf{A} \notin \Delta$. Since $\Gamma \simeq \Sigma$ and $\Delta \simeq \Sigma$, also $\Gamma \simeq \Delta$. Therefore, $\boxdot\mathbf{A} \notin \Gamma$.

Conversely, suppose that $\boxdot\mathbf{A} \notin \Gamma$. By an argument familiar from modal logic (and used in the proof of Lemma 1 above) it is readily seen that the set

$$\Omega = \{\boxdot\mathbf{B}\colon \boxdot\mathbf{B} \in \Gamma\} \cup \{\neg\boxdot\mathbf{C}\colon \boxdot\mathbf{C} \notin \Gamma\} \cup \{\neg\mathbf{A}\}$$

is consistent; to show this, (S5) and (Necessitation) are essential. Lindenbaum's Lemma guarantees the existence of some maximal consistent set $\Delta \supseteq \Omega$. By construction, $\Delta \simeq \Gamma$, and so $\Delta \simeq \Sigma$. Therefore, by Lemma 2, $\Delta$ fits some $\langle T, y\rangle \in \mathfrak{S}_\Sigma$. By the induction hypothesis, the fact that $\mathbf{A} \notin \Delta$ implies that $\langle T, y\rangle \not\models \mathbf{A}$. Hence $\langle S, x\rangle \not\models \boxdot\mathbf{A}$, as we wanted. ∎

In effect the last lemma establishes the strong completeness theorem for M.I.L. stated at the end of Section 2. For let $\Xi$ be any consistent set of formulas. By Lindenbaum's Lemma there is some maximal consistent extension $\Xi^*$ of $\Xi$. By Lemma 2 there is some $\langle S, x\rangle \in \mathfrak{S}_{\Xi^*}$, such that $\Xi^*$ fits $\langle S, x\rangle$ over itself. Consequently, for every $\mathbf{A} \in \Xi^*$, $\langle S, x\rangle \models^{\mathfrak{M}_{\Xi^*}} \mathbf{A}$.

## 4. IS THERE A LOGIC OF INTENTION?

A conspicuous feature of the axiomatization of M.I.L. is that the operator Int plays such a limited rôle in it. Similarly conspicuous is the absence from the formal semantics of any condition on the set $S$ of intentions. Is there no logic of intention? In [2] it was boldly asserted that there is. There the axioms included all instances of the schema

(I1) $\qquad (\mathsf{Int}\,\alpha \wedge \mathsf{Int}\,\beta) \to \alpha = \beta,$

and a semantic condition was adopted which amounted to the following:

(i1) for every possible action $\langle S, x\rangle$ in an intentional action structure, $S$ is either a singleton or else empty; that is, either $S = \{A\}$, for some $A \subseteq U$, or else $S = \emptyset$.

This might seem like an appropriating modelling if agents always had just one intention when acting. However, Example 3 of Section 1 shows that this need not be the case. It will be useful at this point to review that example. In this paper we have in effect identified an intention with an event – the idea is that an intention can always be depicted as an intention to bring about a certain event. On this identification the following account of Example 3 can be made. The agent has decided to give X a birthday present. It seems perfectly acceptable to say that he has formed an intention to give X a birthday present even before he has made up his mind exactly how to implement his decision. This first intention is identified here with the event that the agent gives a birthday present to X. The agent then goes on to form other intentions: on our identification they are more specific or (if you think of their set theoretic representation) less inclusive events. In sum, the deliberation that the agent goes through can be described as a *deliberation walk* in the space of possible events – not outcomes! – which proceeds from more inclusive to less inclusive events and stops when an operational intention is reached. If one accepts this way of viewing things, then one will want a weaker axiom schema than (I1):

(I2) $(\mathsf{Int}\,\alpha \wedge \mathsf{Int}\,\beta) \rightarrow (\alpha \leq \beta \vee \beta \leq \alpha)$.

The corresponding semantic condition would be

(i2) for every possible action $\langle S, x\rangle$ in an intentional action structure, if $A, B \in S$, then either $A \subseteq B$ or $B \subseteq A$.

Example 3 has two features which make it less than general. One is the fact that the deliberation walk stops as soon as an operational intention has been reached. Let us call an intention *operative* if it is the intention on which the agent finally acts. Every operative intention is operational, but there seems to be no reason to assume that the converse must hold. According to the theory presented here there is a unique operative intention in every case of simple intentional action. In a deliberation walk the operative intention would presumably be the one to be added last in time to the agent's set of intentions. Whether it must also be the most specific – most inclusive – intention would remain to be discussed; cf. below.

The other feature that makes Example 3 less than general is that the deliberation walk yields a set of intentions that is linear with respect to the superset relation. But there is no logical reason why a deliberation walk must be linear, and it may well be argued that there are examples to the contrary. Changing Example 3 slightly, let us assume that the agent first decides to give X a birthday present, and then decides to give X a book and also decides that the gift must not cost more than \$10. According to this revamped example, the agent has three intentions: $A$ = "to give X a birthday present", $B$ = "to give X a book", $C$ = "to spend at most \$10 on X". The set $\{A, B, C\}$ is not linear under the superset relation, as $B$ and $C$ are incomparable. While this is no objection to the set $\{A, B, C\}$ as a set of intuitions, it is in a sense objectionable that it is $\cap$-incomplete: it does not contain the further possible intention $B \cap C$ = "to give X a book costing at most \$10". The sense in which it is objectionable is this: it would be odd if the agent were to end his deliberation walk without taking at least one more step and add $B \cap C$ to his set of intentions. He might of course give up his intention to give X a birthday present; this would not be odd, for people often change their mind. Or he might be prevented from carrying his deliberation to a conclusion so that his intentions $A, B, C$ are left floating, as it were. But under normal conditions we would certainly expect an agent who holds both intentions $B$ and $C$ to work his way to intention $B \cap C$ as well: it would be *unreasonable* of him not to do so. This is an interesting point, and we shall return to it below. (Even this step may not be enough, of course: if $B \cap C$ is not operational, the agent will want to continue his deliberation walk.) If this is accepted, then (I2) is too strong as an axiom schema. On the other hand we should need the weaker

(I3) $\quad (\mathsf{Int}\,\alpha \wedge \mathsf{Int}\,\beta) \rightarrow \mathsf{Int}\,(\alpha \frown \beta).$

Similarly, the semantic condition (i2) would be too strong and should be replaced by the weaker condition

(i3) for every possible action $\langle S, x\rangle$ in an intentional action structure, if $A, B \in S$, then $A \cap B \subseteq S$.

One may ask if (I3) and (i3) are not also too strong, and one might go on and try to think of ways to weaken them. For example, if in the revised Example 3 the agent were to acquire a non-empty intention strictly more specific than $B \cap C$, still without acquiring $B \cap C$ itself, then the charge of

unreasonableness would itself be unreasonable (or would we now say that $B \cap C$ is somehow implicit in his set of intentions?). That is to say, it may be worth considering the even weaker condition

(i4) for every possible action $\langle S, x\rangle$ in an intentional action structure, if $A, B \in S$, then $C \subseteq A \cap B$, for some $C \in S$.

However, this condition is more difficult to characterize syntactically, and we will not pursue this discussion further here.

## 5. THREE OBJECTIONS

We will now consider some objections to the theory presented in this paper which serve to point out some important limitations to it. The first objection was raised in conversation by David Lewis (although Lewis must not be held responsible for any detail in the following discussion; the two examples below are the author's). This objection is directed against (I3) and is based on instances in which the agent's intentions are based on error: where he acts on mistaken belief and therefore, without realizing it, attempts to do the impossible. This is a favourite ingredient in fiction and legend. Perhaps Oedipus provides the most famous example of this kind. With his background Oedipus must have been very particular in what men he would kill and what women he would marry. Laius he killed in hot blood, and so it may be argued that he momentarily forgot his normal caution and that his standing intention "to avoid killing my father" for a few fateful seconds failed to be an intention of his. But his marriage to Jocasta must have been a carefully considered *mariage de convenance*, and so among Oedipus' intentions must have been $A$ = "to marry this woman" as well as the standing intention $B$ = "to avoid marrying my mother". As Oedipus found out later, those two intentions were incompatible in the sense that they could not both be realized: $A \cap B$ is impossible, that is, $A \cap B = \emptyset$. But if (I3) is correct, then $A \cap B$ is an intention of Oedipus, and so we must conclude that the impossible is also an intention of Oedipus. Now we can see what the objection is, for $\emptyset$ was surely never an intention of Oedipus.

Before trying to deal with this objection, let us first have another variation on the same theme. Not only (I3) but M.I.L. itself is too strong to pass as a logic of intention, it may be argued, for it contains the theorem

(EI), whence $\alpha = \beta \vdash \mathsf{Int}\,\alpha \leftrightarrow \mathsf{Int}\,\beta$. The event that Electra greets the stranger in front of her is the same as the event that she greets Orestes, since unbeknownst to her Orestes is the stranger in front of her and we have defined events in terms of outcomes. Yet her intention $C$ = "to greet the stranger in front of me" cannot be identified with the possible intention $D$ = "to greet Orestes" – she had the former but not the latter.

Yes, this is true: we need a more sophisticated theory to deal with Oedipus and Electra. In possible worlds language, Oedipus mistook some possible world $w_1$ and Electra some possible world $w_2$ for the actual world $w_0$. In $w_1$ the intentions "to marry this woman" and "to avoid marrying my mother" relativized to Oedipus, *are* compatible, while in $w_2$ the stranger in front of Electra *is not* Orestes. In order to handle examples of this kind we evidently need a richer theory than that developed so far. On the syntactic side we might enrich the object language with an operator $\Box$ (of category $\mathfrak{f}^{\mathfrak{f}}$) expressing universal necessity (truth in all possible worlds). On the semantic side we would add possible worlds. Moreover, we would no longer base intentional action structures on single outcome spaces but on *sets* of outcome spaces, say – this may turn out to be too simple – one for each possible world. Events would now be something like functions – total, say, for simplicity's sake – from possible worlds to sets of outcomes, and if we would continue to identify intention with intention to bring about an event, then intentions too would become functions from possible worlds to sets of outcomes. Giving up (EI) we would still insist on a weaker principle, viz.,

$$(\mathrm{EI}') \qquad \Box(\alpha = \beta) \vdash \mathsf{Int}\,\alpha \leftrightarrow \mathsf{Int}\,\beta.$$

Within a theory thus enriched it would seem possible to accommodate the Electra example. Even though in our world the possible intentions $C$ and $D$ come to the same thing, as $C(w_0) = D(w_0)$, yet in $w_2$ they do not, for $C(w_2) \neq \emptyset$ while $D(w_2) = \emptyset$. Thus $C \neq D$, and so Electra may well have had the intention $C$ without having had the intention $D$.

In a similar way it would seem possible to accommodate also the Oedipus example within the enriched theory. If "this woman" is thought of rigidly, we can account for the tragedy in set theoretic terms: what explains Oedipus' behaviour is that $\emptyset \neq A(w_1) \subset B(w_1)$, and what brings about his fall is that $A(w_0) \cap B(w_0) = \emptyset$. It seem reasonable to define $A \cap B$ as the function $\phi$ such that, for all possible worlds $w$, $\phi(w) = A(w) \cap B(w)$.

Furthermore, let 0 be the intention that is absurd in the sense that it picks out the empty set of outcomes in every possible world. Then $A \cap B \neq 0$, for $(A \cap B)(w_1) \neq \emptyset$, and so it is possible that Oedipus should have had $A \cap B$ but not 0 among his intentions.

Problems attach to the working out of such an enriched semantics which are passed over in silence for now. But it seems that even the preceding sketch would support the contention that Oedipus and Electra type examples are really arguments against (E1) and not against (I3).

The second objection is directed against M.I.L. itself. It was made by an anonymous referee who pointed out, in his report on the penultimate version of this paper, that in M.I.L.

$$\vdash \mathsf{Int}\,\alpha \leftrightarrow \mathsf{Int}\,(\alpha \frown (\beta \smile \bar{\beta})).$$

This, he argued, is unsatisfactory: "It doesn't seem plausible that if I intend to close the door, then I intend to close the door and to visit Japan or not." A first reply to this objection is to emphasize that the following are four different formulas:

(i) $\mathsf{Int}\,(\alpha \frown (\beta \smile \bar{\beta}))$,

(ii) $\mathsf{Int}\,\alpha \wedge \mathsf{Int}\,(\beta \smile \bar{\beta})$,

(iii) $\mathsf{Int}\,\alpha \wedge (\mathsf{Int}\,\beta \vee \mathsf{Int}\,\bar{\beta})$,

(iv) $\mathsf{Int}\,\alpha \wedge (\mathsf{Int}\,\beta \vee \neg\mathsf{Int}\,\beta)$.

In the recommended jargon of Section 1 they read, "$\alpha \frown (\beta \smile \bar{\beta})$ is intended by the agent", "$\alpha$ and $\beta \smile \bar{\beta}$ are both intended by the agent", "$\alpha$ and either $\beta$ or $\bar{\beta}$ are both intended by the agent", and "$\alpha$ is intended by the agent, and $\beta$ is either intended or not". The referee's objection depends on two circumstances, that $\mathsf{Int}\,\alpha$ is logically equivalent to (i) and (iv), but not to (ii) or (iii), and that an English sentence of type "I intend to close the door and visit Japan or not" may perhaps be taken to be ambiguous between (i), (ii), (iii), and (iv). Thus, according to this line of defence, the air of paradox will disappear if proper care is taken when one translates between English and formulas.

The objectionable feature is deeply grounded in the semantics presented here. It depends on two important principles that have guided our analysis: that intention is to be understood as intention to bring about an event, and

that an event is determined by its instantiating outcomes. If one would like to develop a system escaping the referee's objection, at least one of these principles will have to be given up. Notice that also the enriched system sketched in preceding paragraphs fails to escape it: in every possible world the functions $\|\alpha\|$ and $\|\alpha \frown (\beta \smile \bar{\beta})\|$ single out the same set of outcomes, and so they are (almost) the same function. How to develop an interesting system without this property is not clear to the author. For one effort to solve what seems to be virtually the same problem, see von Wright [3]; but even there the objection survives in a weakened form.

However, before defeat is conceded one may also try a second, more aggressive line of defence: is the referee's objection really an objection? The author, for his part, is inclined to accept the referee's "paradox". The logic of action presented here does at least not seem to be in a worse position than possible worlds type intensional logic generally. In most epistemic logics we have

$$\vdash \mathbf{KA} \leftrightarrow \mathbf{K}(\mathbf{A} \wedge (\mathbf{B} \vee \neg \mathbf{B})),$$

and it may appear odd that I cannot know that the door is closed without knowing that the door is closed and that Japan beat Sweden in football in the 1936 Olympics or not. A related and better known version of this difficulty is the paradox of the logically omniscient subject: in most epistemic logics, an agent who knows one logical truth knows them all. Should one wish to develop an epistemic logic which does not suffer from this defect, then one should first try to make it clear what limitations one should like to attribute to the knowing subject. As there are limitations of different sorts, different logics may ensue. But even so there will always be a residue of paradox: just as the logically omniscient subject suffers from perfection, so any logically limited subject is likely to suffer from perfection within his limits. What is meant by this assertion will be made clear in our discussion of the third objection, to which we now turn.

The third objection is that not only (I3) but any proper extension of M.I.L. will be too strong if one has the logic of intentions of real people in mind; for real people are free to do all sorts of things, and there are real life situations that cannot be modelled in even the weakest of the three extensions of M.I.L. that we have discussed in this section. According to this objection it is enough to consider Example 3 as revamped above. It is all right for us to say, as we did, that it would be unreasonable of the agent to

intend $B$ and intend $C$, yet not intend $B \cap C$. But however loudly we say this, it is nevertheless possible to imagine exactly this kind of situation. That is to say, the set $\{\mathsf{Int}\ \pi_0, \mathsf{Int}\ \pi_1, \neg\mathsf{Int}\ (\pi_0 \frown \pi_1)\}$, where $\pi_0$ and $\pi_1$ are event letters, would seem to be logically consistent, and thus rule out (I3).

This objection is of a different kind from the previous two, and much more far-reaching. The present objection could hardly be met by strengthening the expressive power of our theory. On the other hand, if the objection is accepted, trying to meet it would lead to a logic of intention so weak as to be void of content. The dilemma is solved by reflecting on the theoretical status of our theory. The logic of intention and action is not different from many other kinds of philosophical logic. Many of the difficulties and unsolved problems of intensional logic (with an 's') are going to be problems also for our intentional logic (with a 't'). Among other things there is the problem of applicability. Lewis' system S4 and related systems are not much use in a study of what people actually know or believe. For example, knowledge or belief operators cannot be depended upon to distribute over conjunction where actual knowers or believers are concerned. But this does not necessarily mean that the usual systems of epistemic and doxastic logic are uninteresting! We encounter the same problem in game theory and decision theory, disciplines whose claim to interest does not lie in their ability to describe actual behaviour. In other words, examples that tell against a certain theory seen as a descriptive modelling need not tell against that very same theory seen as a normative modelling. Therefore, notwithstanding the last objection, if one is asking for a logic of *rational action*, the extension of M.I.L. by (13) looks like a possible candidate.

*University of Auckland*

## REFERENCES

[1] Segerberg, Krister, Applying modal logic, *Studia Logica* 39 (1980), 275–95.
[2] Segerberg, Krister, Action-games, *Acta Philosophica Fennica*, to appear.
[3] von Wright, Georg Henrik, Deontic logic revisited, *Rechtstheorie* 4 (1973), 37–46.