First Steps in the *stit*-Logic Analysis of Intentional Action

Jan Broersen

Abstract

We study intentional action in *stit*-logic. The formal logic study of intentional action appears to be new, since most logical studies of intention concern intention as a mental state. We model an intentional action as an action that possibly deviates from the actual action conducted by an agent. First, an actual action may deviate from an intended action because the agent is not able to carry out the intended action, and the actual action is used as a *means* to achieve the goal of the intended action. This explains differences such as the one between the actions 'murder' and 'manslaughter', and problems like the 'dentist's' scenario of Cohen and Levesque. Second, an actual action may deviate from an intended action because the agent's environment behaves unexpectedly and the result of an action is not the one envisaged by the agent. So, the action is unexpectedly *unsuccessful*. We show how to deal with the distinction between successful and non-successful action by weakening the notion of 'knowingly doing' to its 'belief' equivalent.

1 Introduction

This paper studies intention as a mode of acting. This is quite different from studying intention as one of the elements of mental states of agents. Within the computer science community working on logical approaches to AI, the best-known paper on the latter subject is the one by Cohen and Levesque [7]. The difference in subject between the present work and that work is best explained by of Cohen and Levesque themselves [7](p 216):

> Most philosophical analysis has examined the relationship between an agent's doing something intentionally and that agent's having a present-directed intention. Recently, Bratman [2] has argued that intending to do something (or having an intention) and doing something intentionally are not the same phenomenon, and that the former is more concerned with the coordination of an agent's plans. We agree, and in this paper we concentrate primarily on future-directed intentions. Hereafter, the term "intention" will be used in that sense only.

In this paper we study the interpretation of intention explicitly excluded by Cohen and Levesque: "intentionally doing". The difference is parallelled by differences in the

formal apparatuses to study the notions. Cohen and Levesque use a first-order logic where action is represented in the same way as in Dynamic Logic [21, 14]. More precisely: to talk about action they use a translation of propositional dynamic logic into their first-order language. But, they do not have a construct in their language representing acting as such. Like in dynamic logic, in the formalism there is no object level construct for expressing, for instance, "agent A writes a paper". As in Dynamic Logic, expressivity is limited to conditional assertions like "if action a would be executed, it would have as an effect that a paper is written" and non-conditional assertions like "action a is executable". The reason that Cohen and Levesque do not need an operator for action is that they do not study intentional action, but intention as a mental state. For our study of intention as a mode of acting, we use *stit*-logic. *stit*-logic does enable us to talk about action directly. For the present study, another advantage of using *stit*-logic rather than dynamic logic is that it is still unclear how to express properties of agency in dynamic logic (examples of such properties are: refraining, deliberate action, independence of agency, regularity, etc.).

Having pointed out the difference with the work of Cohen and Levesque, we want to stress that there are issues that arise under both interpretations of 'intention'. In particular, one of the central issues in the work of Cohen and Levesque is that intention is not closed under side effects of action (the well-known denstist's example). In our framework we will analyze and solve the same problem in the context of intentional action.

Intentional action is studied in philosophy since Anscombe's seminal work on the subject [1]. But our main motivation for the present work comes from the literature on law and deontic logic. As is well known, for a judge deciding on a verdict, there is a lot of difference between murder, manslaughter, homicide, killing in self-defense, etc. Yet, all these acts concern the same *physical* event: that of causing someone's death. The difference is in the mode of acting, that is, in the mental state by which the agent's act is accompanied at the time of conduct (the legal literature speaks of 'showing concurrence').

In criminal law, the different modes of acting correspond with different categories of culpability. And it is the judge's task to assess to which category a case belongs. Of course, different law systems have different categories. The current North American system works with the following modes, in decreasing order of culpability (as taken from [9]):

- **Purposefully** - the actor has the "conscious object" of engaging in conduct and believes and hopes that the attendant circumstances exist.
- **Knowingly** - the actor is certain that his conduct will lead to the result.
- **Recklessly** - the actor is aware that the attendant circumstances exist, but nevertheless engages in the conduct that a "law-abiding person" would have refrained from.
- **Negligently** - the actor is unaware of the attendant circumstances and the consequences of his conduct, but a "reasonable person" would have been aware
- **Strict liability** - the actor engaged in conduct and his mental state is irrelevant

In this paper we will be only concerned with the first two categories. We aim to formalize the distinctions between the other categories in the journal version of [3].

The first category, the one of acts committed *purposefully*, is about acts that are instrumental in reaching an agent's malicious *goal*. So, this is the category of intentional action. The second category is not directly about an agent's intentions, aims or goals, but only about the condition whether or not an agent knows what it is doing.

The plan of this paper is as follows. First, in section 2 we present the base formalism in which we perform our analysis: XSTIT [4]. Then, in section 4 we discuss the notion of 'knowingly doing'. For this notion we will formulate new properties not given in [3]. Then, in section 5 we present our view on the notion of intentionally doing, and discuss the relation and difference with knowingly doing. In section 6 we observe that intentionally doing as defined in section 5 does not leave room for intentional action being non-successful. We show how to adapt the properties to allow for non-successful action. In particular, we will weaken the notion if 'knowingly doing' to its belief equivalent. Finally section 7 discusses future work and conclusions.

2 A group *stit*-logic affecting 'next' states: XSTIT

The logic XSTIT was first investigated in [4]. We also used the almost identical name 'X-STIT' in [5], but there the 'X' is separated from the acronym 'STIT', which refers to the fact that that paper's classical *instantaneous stit* logic is extended with a next operator, while in XSTIT effectivity of *stit*-operators itself refers to next states. That is not the only difference with the *stit*-logic(s) in [5]. In particular, XSTIT drops some of the axioms in [5], adds several new ones, and is complete. Also we use a two dimensional semantics, closer to the *stit*-semantics in the philosophical literature. Because the stit-operators of XSTIT refer to next states, we avoid that the logic is undecidable and not finitely axiomatizable [15]; XSTIT is canonical.

Besides the usual propositional connectives, the syntax of XSTIT comprises an operator $\Box\varphi$ for historical necessity, which plays the same role as the well-known path quantifiers in logics such as CTL and CTL* [10], and an operator $[A\ \mathsf{xstit}]\varphi$ for 'agents A jointly see to it that φ in the next state'. Our *stit*-operator concerns, what game-theorists call, 'one-shot' actions.

$$\varphi \quad := \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [A\ \mathsf{xstit}]\varphi$$

$$X\varphi \quad \equiv_{def} \quad [Ags\ \mathsf{xstit}]\varphi$$

An XSTIT-frame is a tuple $\langle S, H, R_\Box, \{R_A \mid A \subseteq Ags\}\rangle$ such that:

- S is a non-empty set of states. Elements of S are denoted s, s', etc.
- H is a non-empty set of histories. Histories are sets of states. Elements of H are denoted h, h', etc.
- Structured worlds are tuples $\langle s, h\rangle$, with $s \in S$ and $h \in H$ and $s \in h$.
- $R_\Box$ is a 'historical necessity' relation over structured worlds such that $\langle h, s\rangle R_\Box \langle h', s'\rangle$ if and only if $s = s'$

- The R_A are 'effectivity' relations over structured worlds obeying appropriate first-order frame conditions (depending on the Sahlqvist axioms adopted)

We do not give the exact first-order frame conditions here (but see [4]). The reason is that our framework, unlike other *stit*-frameworks, has a standard modal semantics and the axioms are all within the Sahlqvist class. This means they correspond with first-order frame conditions, that can for instance easily be found using the algorithm SQEMA [8].

Validity $\mathcal{M}, \langle s, h\rangle \models \varphi$, of a formula φ in a history/state pair $\langle s, h\rangle$ of a model $\mathcal{M} = \langle S, H, R_\Box, \{R_A \mid A \subseteq Ags\}, \pi\rangle$ is defined as:

$$\begin{array}{lcl} \mathcal{M}, \langle s, h\rangle \models \Box\varphi & \Leftrightarrow & \langle s, h\rangle R_\Box \langle s', h'\rangle \text{ implies that } \mathcal{M}, \langle s', h'\rangle \models \varphi \\ \mathcal{M}, \langle s, h\rangle \models [A\ \mathsf{xstit}]\varphi & \Leftrightarrow & \langle s, h\rangle R_A \langle s', h'\rangle \text{ implies that } \mathcal{M}, \langle s', h'\rangle \models \varphi \end{array}$$

Satisfiability, validity on a frame and general validity are defined as usual. The following axiom schemas, in combination with a standard axiomatization for propositional logic, and the standard rules (like necessitation) for the normal modal operators, define a Hilbert system for XSTIT:

	S5 for $\Box$
	KD for each $[A\ \mathsf{xstit}]$
(Det)	$\neg X \neg\varphi \rightarrow X\varphi$
(C-Mon)	$[A\ \mathsf{xstit}]\varphi \rightarrow [A \cup B\ \mathsf{xstit}]\varphi$
($\emptyset \Rightarrow$ Sett)	$[\emptyset\ \mathsf{xstit}]\varphi \rightarrow \Box X\varphi$
(X-Eff)	$\Box X\varphi \rightarrow [A\ \mathsf{xstit}]\varphi$
(NCUH)	$[A\ \mathsf{xstit}]\varphi \rightarrow X\Box\varphi$
(Indep-G)	$\Diamond[A\ \mathsf{xstit}]\varphi \wedge \Diamond[B\ \mathsf{xstit}]\psi \rightarrow \Diamond([A\ \mathsf{xstit}]\varphi \wedge [B\ \mathsf{xstit}]\psi)$ for $A \cap B = \emptyset$

It is easy to check that all the axioms we give are within the Sahlqvist class, which means they correspond to first order conditions on the frames and are complete with respect to these frames ([4]).

Pauly's Coalition logic (CL) [20] is a logic of ability that is very closely related to *stit*-formalisms. XSTIT contains CL as a fragment ([4]).

3 Operators for knowledge and intention

We extend XSTIT with epistemic operators $K_a\varphi$ for knowledge of individual agents a, and operators $[a\ \mathsf{xint}]\varphi$ for agent a intends doing φ.

Herzig and Troquard were the first to consider the addition of knowledge operators to a *stit*-logic [16]. Later on the framework was adapted and extended by Broersen, Herzig and Troquard [5, 6]. The epistemic fragment of the present logic extends our earlier work on epistemic *stit* in several ways. In particular, new properties for the interaction of knowledge and action are proposed. Also the semantics, being two-dimensional, is different from the one in [5]. Finally, the modeled concept is 'knowingly doing', whereas in e.g. [16] the aim is to model 'knowing how'.

Intention operators have been considered in the *stit*-framework by Lorinin and Herzig [11, 18] and by Semmling and Wansing [22]. However, in both these works, like

in the work of Cohen and Levesque, the emphasis is on intention as a mental state, and not on intention as a mode of acting. Since we study intention as a mode of acting, our intention operator $[a\ \mathsf{xint}]\varphi$ deliberately has the same appearance as the *stit*-operator $[A\ \mathsf{xstit}]\varphi$. The basic idea is that intended actions are actions in an idealized, mental sense. Although they are not 'physical', they obey properties of action. The main point is that they should not be identified with the actual (physical) actions carried out by the agent. So, an intended action is an action, but in another, idealized view on the possible histories selected by the action. The 'x' in the notation $[A\ \mathsf{xstit}]\varphi$ refers to the fact that also the effects of intended actions are realized in 'next' states.

Formally, we extend XSTIT's syntax as follows.

$$\varphi \ldots \quad := \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [A\ \mathsf{xstit}]\varphi \mid K_a\varphi \mid [a\ \mathsf{xint}]\varphi$$

Note that the stit-operators concern groups of agents, while the knowledge and intention operators concern individual agents. In this paper we do not want to consider the intricacies of the action versions of group knowledge and group intention.

We extend XSTIT's semantic basis by the following definitions.

A frame is a tuple $\langle S, H, R_\Box, \{R_A \mid A \subseteq Ags\}, \{\sim_a \mid a \in Ags\}, \{I_a \mid a \in Ags\}\rangle$ such that:

- $\langle S, H, R_\Box, \{R_A \mid A \subseteq Ags\}\rangle$ is an XSTIT-frame
- The $\sim_a$ are epistemic equivalence relations over structured worlds obeying appropriate first-order conditions (depending on the Sahlqvist axioms adopted).
- The I_a are intentional effectivity relations over structured worlds obeying appropriate first-order conditions (depending on the Sahlqvist axioms adopted).

The clause for validity of formulas is extended with:

$$\begin{array}{lcl} \mathcal{M}, \langle s, h\rangle \models K_a\varphi & \Leftrightarrow & \langle s, h\rangle \sim_a \langle s', h'\rangle \text{ implies that } \mathcal{M}, \langle s', h'\rangle \models \varphi \\ \mathcal{M}, \langle s, h\rangle \models [a\ \mathsf{xint}]\varphi & \Leftrightarrow & \langle s, h\rangle I_a \langle s', h'\rangle \text{ implies that } \mathcal{M}, \langle s', h'\rangle \models \varphi \end{array}$$

4 Knowingly doing

With the above definitions we can express that agent a *knowingly* sees to it that φ as $K_a[a\ \mathsf{xstit}]\varphi$ [3]. Semantically: an agent knowingly does something if its action 'holds' for all the history/state pairs in the epistemic equivalence set containing the *actual* history/state pair. In [5] we also called this 'conformantly' doing, in analogy with the notion of conformant planning [13], which looks at plans that are successful under incomplete knowledge of the current state.

We now give some new properties for the interaction between knowledge and action, most of which were not given in [3]. The properties are, again, in the Sahlqvist class, and their corresponding first order conditions can thus be added to the semantics to

obtain a complete system.

	All XSTIT axioms
	S5 for each K_a
(K-Contr)	$K_a[A\ \mathsf{xstit}]\varphi \rightarrow K_a[a\ \mathsf{xstit}]\varphi$ for $a \in A$
(NK-Oth)	$K_a[A\ \mathsf{xstit}]\varphi \rightarrow K_a\Box[A\ \mathsf{xstit}]\varphi$ for $a \notin A$
(Rec-Eff)	$K_a[a\ \mathsf{xstit}]\varphi \rightarrow XK_a\varphi$
(Unif-Str)	$\Diamond K_a[a\ \mathsf{xstit}]\varphi \rightarrow K_a\Diamond[a\ \mathsf{xstit}]\varphi$
(K-S)	$K_a\Box\varphi \rightarrow \Box K_a\varphi$

The axioms express intuitive properties for the interactions of knowledge and action. $(K-Contr)$ expresses that an agent can only know that a larger group that it is part of ensures something if the agent ensures that thing himself. $(NK-Oth)$ expresses that agents have no means whatsoever to know what it is that groups they are *not* part of bring about concurrently. $(Rec-Eff)$ expresses that if agents knowingly see to something, than they know that something is the case in the resulting state (also discussed in [3]). $(Unif-Str)$ expresses that if an agent can knowingly see to something, it knows seeing to that something is one of its causal capacities. For instance: the fact that I can knowingly break the cup by throwing it on the floor implies that I know to have the causal power to break the cup. Finally, $(K-S)$ expresses that knowing that something is settled implies that it is settled that it is known.

That knowledge has an entirely different character here than in most systems with epistemic operators, is maybe best explained through the notion of 'moment determinacy' [17]. Semantically, moment determinacy of an operator M is defined by the condition that the truth value of M is independent of the history h in structured worlds $\langle s, h\rangle$. Syntactically, moment determinacy can be defined as follows: M is moment determinate if $M\varphi \rightarrow \Box M\varphi$ is valid. An example of a moment determinate modality is 'unconditional obligation' (however, see [23] for a different opinion on the moment determinacy of obligation). In general it is assumed that what one is unconditionally obliged to do does not depend on what one does. Of course the exception is when one considers obligations that are explicitly conditional on what an agent does (if you drive a car, you need to carry your license; if you kill, you have to kill gently [12]).

Now, in the present framework, knowledge is not moment determinate. We cannot conclude to $K_a\varphi \rightarrow \Box K_a\varphi$, because that does not hold for the substitution $[[a\ \mathsf{xstit}]\psi/\varphi]$. And this seems right: an agent's knowledge should not only depend on the moment of consideration. If we can assume that an agent knows what it does when it chooses something, what it knows depends on what it chooses to do, and not only on the state.

5 Intentionally doing

First we give a number of axioms that reflect our idea that intentionally doing is doing in a more abstract sense. First of all, intentional action is consistent, that is, it cannot be consistent that an agent intentionally sees to it that φ and at the same time intentionally sees to it that $\neg\varphi$. This gives us KD for the individual intentional action modalities. Note that the *xstit*-modality is also KD. Second, also for intentional

action we adopt the independence of agency axiom: imagining that intended actions of agents are not independent is even harder than imagining that choices of agents are independent (the *Indep-G* axiom of section 2).

KD for each $[a\ \mathsf{xint}]$

(Indep-I) $\Diamond[a\ \mathsf{xint}]\varphi \wedge \Diamond[b\ \mathsf{xint}]\psi \rightarrow \Diamond([a\ \mathsf{xint}]\varphi \wedge [b\ \mathsf{xint}]\psi)$

Note that intentional action is not moment determinate. Ones more this emphasizes that we do not model intention as a mental state.

Now we turn to the very important interaction of intentional action with knowledge. First, it seems correct to assume that intentional action has its result among the states the agent *knows* to be possible next states. In XSTIT (section 2) the *X-Eff* axiom expresses for physical action that they take effect in next states. Here, for intentional action, we need a variant of this axiom involving knowledge:

(X-Eff-I) $K_a \Box X\varphi \rightarrow [a\ \mathsf{xint}]\varphi$

So, if the agent knows that φ is settled for the next state, which means he cannot do anything about it, he cannot but intend that φ holds next. Of course this is strange in a 'deliberate' reading of intentional action. But here we use versions of the operators that are closed under logical consequence (which, for instance, gives us $[a\ \mathsf{xint}]\top$ as a theorem), and we do not consider the weaker *deliberate* versions of the operators not closed under logical consequence. As is well-known they can easily be defined as syntactic abbreviations.

Now we go to the second interaction with knowingly doing. A crucial property of intended action seems to me that an agent only performs an intended action if that same agent performs that action knowingly. If I send an email, and by doing that I do not *knowingly* cause a server to break down, I clearly do not *intentionally* bring down the server by sending that email. Within the present framework, we can capture this property of intentionally doing by the following Sahlqvist axiom.

(I-K) $[a\ \mathsf{xint}]\varphi \rightarrow K_a[a\ \mathsf{xstit}]\varphi$

This axiom gives rise to three important observations. The first is that clearly, we do not want to impose the other direction of the implication as an axiom. The axiom *(I-K)* explains why in law, the category of 'purposefully' conducted acts (see section 1) is higher in hierarchy of culpability than the category of 'knowingly' conducted acts: the former implies the latter, but not the other way around. An agent killing in self-defense, kills knowingly, but does not kill intentionally.

The second, closely related point is that the axiom *(I-K)* explains that intention is not closed under what Cohen and Levesque [7] call 'side effects' of action. A side effect is an effect of an action that is not intended. And the issue is that in our formalization, intentional action should thus not be closed under these side effects.

An action with side effects deviates from the intended action in that it has extra effects. So, the relation between the two acts is one of action implication / action subsumption. In our formalism we can express action implication / action subsumption as follows. The *stit* formula $\Box([a\ \mathsf{xstit}]\varphi \rightarrow [a\ \mathsf{xstit}]\psi)$ expresses that for agent a,

presently, doing φ is a way of doing ψ. More precisely, it says that any *way* of doing φ is a *way* of doing ψ. So presently, the coalition cannot do φ without doing ψ. One can also say that A can do φ as a 'means' to do ψ. After all, there might be more choices that ensure ψ and doing ψ by doing φ is thus possibly only one way of doing ψ. Note that the formula does not say that agents A are able to do either φ or ψ; the formula only expresses a relation between the two actions.

Let us go back to the examples already addressed in passing. For the dentist's example, we can for instance say that $[a\ \mathsf{xstit}](d \wedge p)$ = "visiting the dentist and have pain" and $[a\ \mathsf{xstit}]d$ = "visiting the dentist". Clearly $[a\ \mathsf{xstit}](d \wedge p)$ is way of doing $[a\ \mathsf{xstit}]d$ and we have as a logical fact that $\Box([a\ \mathsf{xstit}](d \wedge p) \rightarrow [a\ \mathsf{xstit}]d)$. Now assume the agent intentionally visits the dentist but has no other way of doing that than by going to the dentist and have pain. So, at that moment in time for agent a the situation is such that there is a *causal* relation from $[a\ \mathsf{xstit}]d$ to $[a\ \mathsf{xstit}](d \wedge p)$, that is, in the direction opposite to the direction of the *logical* relation expressed by $[a\ \mathsf{xstit}](d \wedge p) \rightarrow [a\ \mathsf{xstit}]d$. Formalizing the situation of the agent, we come to the set of formulas $Th = \{[a\ \mathsf{xint}]d,\ K_a[a\ \mathsf{xstit}](d \wedge p),\ \neg\Diamond K_a[a\ \mathsf{xstit}](d \wedge \neg p)\}$, that is, (1) the agent intentionally visits the dentist, (2) knowingly visits the dentist in a way that causes him pain, and (3) does not know a way of visiting the dentist without having pain. Clearly we derive that the agent knowingly sees to it that it has pain ($Th \vdash K_a[a\ \mathsf{xstit}]p$). But, we cannot derive that the agent intentionally sees to it that it has pain ($Th \nvdash [a\ \mathsf{xint}]p$); counter models can be constructed.

A similar structure emerges in the example concerning the difference between killing intentionally and killing out of self-defense. A person stabs another person out of self-defense. His intention is not to kill the other person. His intention is to protect himself. But at that particular point in time, the only way to protect himself is by killing the other person. Now, killing the other person is clearly a *consequence* of the agent having the intention to protect himself. But it is not a logical consequence; it is a causal consequence at that particular moment in time. Assume $[a\ \mathsf{xstit}](d \wedge k)$ = "defending and killing" and $[a\ \mathsf{xstit}]d$ = "defending". Now the situation is modeled by $Th = \{[a\ \mathsf{xint}]d,\ K_a[a\ \mathsf{xstit}](d \wedge k),\ \neg\Diamond K_a[a\ \mathsf{xstit}](d \wedge \neg k)\}$, that is, (1) the agent intentionally defends itself, (2) knowingly defends itself in a way that kills another person, and (3) does not know a way of defending itself other than by killing the person. We derive that the agent knowingly sees to it that the person is killed ($Th \vdash K_a[a\ \mathsf{xstit}]k$), but not that the agent intentionally sees to it that the person is killed ($Th \nvdash [a\ \mathsf{xint}]k$).

For the third issue raised by the axiom *(I-K)*, we go to a new section.

6 Non-successful action

Axiom *(I-K)* ensures that an intended action is also a knowingly performed action. Knowingly performed actions are successful actions in the sense that the actual history-state pair is among the pairs in the epistemic equivalence class (game theorists would say: 'the information set'). Axiomatically, we have that from *(I-K)* $[a\ \mathsf{xint}]\varphi \rightarrow K_a[a\ \mathsf{xstit}]\varphi$ and the veridicality of knowledge we derive that $[a\ \mathsf{xint}]\varphi \rightarrow [a\ \mathsf{xstit}]\varphi$. Then with axiom *(NCUH)* we derive that $[a\ \mathsf{xint}]\varphi \rightarrow X\Box\varphi$. Finally, with standard

normal modal reasoning, we arrive at $[a\ \mathsf{xint}]\varphi \rightarrow X\varphi$. Also this axiom says that intentional action is successful: what an agent intentionally does is also what happens.

But, for intentional action this is often simply not the case. What we intentionally do, is not necessarily what happens. For instance, the environment may behave unexpectedly, causing the the actual action to be completely different than the intended action. It can even be the case that we intentionally perform an action and achieve the opposite. For instance, we perform the intentional action of securing a precious vase that is too close to the edge of a table, and by doing so, we cause it to fall on the ground.

The system build so far can be adapted to allow for the fact that intentional action is not successful in a elegant way. What we need to do is to allow for a possible discrepancy between what one thinks one does and what actually happens. So, what we need to do, is to weaken the notion of knowingly doing to its belief equivalent. We do not have a good word for the notion thus resulting; maybe 'believing to do' is the phrase that comes closest.

Let us very briefly present the resulting logic. We change the knowledge operator in a belief operator, resulting in the following syntax.

$$\varphi \ldots \quad := \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [A\ \mathsf{xstit}]\varphi \mid B_a\varphi \mid [a\ \mathsf{xint}]\varphi$$

A frame is now a tuple $\langle S, H, R_\Box, \{R_A \mid A \subseteq Ags\}, \{B_a \mid a \in Ags\}, \{I_a \mid a \in Ags\}\rangle$ such that:

- $\langle S, H, R_\Box, \{R_A \mid A \subseteq Ags\}\rangle$ is an XSTIT-frame
- The B_a are epistemic accessibility relations over structured worlds obeying appropriate first-order conditions (depending on the Sahlqvist axioms adopted).
- The I_a are intentional effectivity relations over structured worlds obeying appropriate first-order conditions (depending on the Sahlqvist axioms adopted).

The new clause for the truth condition of belief is:

$$\mathcal{M}, \langle s, h\rangle \models B_a\varphi \quad \Leftrightarrow \quad \langle s, h\rangle B_a \langle s', h'\rangle \text{ implies that } \mathcal{M}, \langle s', h'\rangle \models \varphi$$

For the axioms, we cannot simply turn the ones for knowingly doing in section 4 into belief equivalents. First of all, of course, we now take KD45 instead of S5 for the individual epistemic operators. But also it seems we have to drop the axioms concerning what other agents do simultaneously. At this point, a deeper investigation is still lacking though. We leave that for further research.

	All XSTIT axioms
	KD45 for each B_a
(Rec-Eff)	$B_a[a\ \mathsf{xstit}]\varphi \rightarrow XB_a\varphi$
(Unif-Str)	$\Diamond B_a[a\ \mathsf{xstit}]\varphi \rightarrow B_a\Diamond[a\ \mathsf{xstit}]\varphi$

Finally, and most importantly, we also need new versions of the axioms concerning the interaction of intention and the epistemic operator. We get:

(X-Eff-I)	$B_a \Box X\varphi \rightarrow [a\ \mathsf{xint}]\varphi$
(I-B)	$[a\ \mathsf{xint}]\varphi \rightarrow B_a[a\ \mathsf{xstit}]\varphi$

We conclude this section with the observation that in the framework with knowledge replaced by belief, we no longer have that intended action is necessarily successful. It is no longer the case that the actual history-state pair is among the pairs that are epistemically accessible. And axiomatically, we no do not have that from *(I-B)* $[a\ \mathsf{xint}]\varphi \rightarrow B_a[a\ \mathsf{xstit}]\varphi$ we derive that $[a\ \mathsf{xint}]\varphi \rightarrow [a\ \mathsf{xstit}]\varphi$, because belief is not like knowledge veridical.

7 Conclusion

We have presented some first steps in the stit-logic analysis of intentional action. We have shown how our formalization avoids that intentional action is closed under side effects: our operator for intentional action is closed under logical consequence, but not under causal consequence. Also we have shown how to represent intentional action that is possibly not successful. We argued that the distinction between successful and non-successful action only makes sense if there can be a distinction between what agents belief to do and what they actually do. If these coincide there is success. If these do not coincide, there is failure. Our suggestion is thus to weaken the notion of 'knowingly doing' to its belief equivalent.

Many issues remain open. One is related to the observation that there are several modes of knowingly doing. For instance, we can knowingly refrain, and we can knowingly allow for a certain outcome. We have not investigated yet the interaction of these modes with intention. For instance, can we intentionally knowingly refrain? Can we intentionally knowingly allow?

Another question is the formalization of the notion of 'attempt'. Attempts are intentional actions that are possibly not successful. Possibly, the difference with the intentional and possibly not successful actions formalized here is that when an agent performs an attempt, it already knows that its action is possibly not successful. Incorporating this aspect might involve having both knowledge and belief operators in the language.

Finally there is the formalization of the notion of 'moral luck' [19, 24]. One way in which an agent can be said to be morally lucky is when the intention of his action is bad, but circumstances cause that the action does not work out as badly as intended. It seems we can represent aspects of this in the present framework.

References

[1] G.E.M. Anscombe. *Intention (2nd edn.)*. Cornell University Press, Ithaca, NY, 1963.

[2] M. Bratman. Two faces of intention. *Philos. Rev.*, (93):375–405, 1984.

[3] J.M. Broersen. A logical analysis of the interaction between 'obligation-to-do' and 'knowingly doing'. In L.W.N. van der Torre and R. van der Meyden, editors, *Proceedings 9th International Workshop on Deontic Logic in Computer Science (DEON'08)*, volume 5076 of *Lecture Notes in Computer Science*, pages 140–154. Springer, 2008.

[4] J.M. Broersen. A complete stit logic for knowledge and action, and some of its applications. In Matteo Baldoni, Tran Cao Son, M. Birna van Riemsdijk, and Michael Winikoff, editors, *Declarative Agent Languages and Technologies VI, 6th International Workshop, DALT 2008, Estoril, Portugal, May 12, 2008, Revised Selected and Invited Papers*, volume 5397 of *Lecture Notes in Computer Science*, pages 47–59, 2009.

[5] J.M. Broersen, A. Herzig, and N. Troquard. A normal simulation of coalition logic and an epistemic extension. In *Proceedings Theoretical Aspects Rationality and Knowledge (TARK XI), Brussels.*

[6] J.M. Broersen, A. Herzig, and N. Troquard. A STIT-extension of ATL. In Michael Fisher, editor, *Proceedings Tenth European Conference on Logics in Artificial Intelligence (JELIA'06)*, volume 4160 of *Lecture Notes in Artificial Intelligence*, pages 69–81. Springer, 2006.

[7] P.R. Cohen and H.J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(3):213–261, 1990.

[8] W. Conradie, V. Goranko, and D. Vakarelov. Algorithmic correspondence and completeness in modal logic I: The core algorithm SQEMA. *Logical Methods in Computer Science*, 2(1):1–26, 2006.

[9] Markus D. Dubber. *Criminal Law: Model Penal Code.* Foundation Press, 2002.

[10] E.A. Emerson. Temporal and modal logic. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science, volume B: Formal Models and Semantics*, chapter 14, pages 996–1072. Elsevier Science, 1990.

[11] Andreas Herzig Cristiano Castelfranchi Emiliano Lorini, Nicolas Troquard. Delegation and mental states. In *Proceedings of Sixth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'07)*, New York, NY, USA, 2007. ACM Press.

[12] J.W. Forrester. Gentle murder, or the adverbial Samaritan. *Journal of Philosophy*, 81(4):193–197, 1984.

[13] Robert P. Goldman and Mark S. Boddy. Expressive planning and explicit knowledge. In *Proceedings of the 3rd International Conference on Artificial Intelligence Planning Systems (AIPS-96)*, pages 110–117. AAAI press, 1996.

[14] D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic.* The MIT Press, 2000.

[15] Andreas Herzig and Francois Schwarzentruber. Properties of logics of individual and group agency. In Carlos Areces and Rob Goldblatt, editors, *Advances in Modal Logic*, volume 7, pages 133–149. College Publications, 2008.

[16] Andreas Herzig and Nicolas Troquard. Knowing How to Play: Uniform Choices in Logics of Agency. In Gerhard Weiss and Peter Stone, editors, *5th International Joint Conference on Autonomous Agents & Multi Agent Systems (AAMAS-06), Hakodate, Japan*, pages 209–216. ACM Press, 8-12 May 2006.

[17] J.F. Horty. *Agency and Deontic Logic*. Oxford University Press, 2001.

[18] E. Lorini and A. Herzig. A logic of intention and attempt. *Synthese*, 163(1):45–77, 2008.

[19] Thomas Nagel. Moral luck. In *Mortal Questions*, pages 24–38. Cambridge University Press, 1979.

[20] Marc Pauly. A modal logic for coalitional power in games. *Journal of Logic and Computation*, 12(1):149–166, 2002.

[21] V.R. Pratt. Semantical considerations on Floyd-Hoare logic. In *Proceedings 17th IEEE Symposium on the Foundations of Computer Science*, pages 109–121. IEEE Computer Society Press, 1976.

[22] Caroline Semmling and Heinrich Wansing. FROM BDI AND stit TO bdi-stit LOGIC. *Logic and Logical Philosophy*, (17):185–207, 2008.

[23] H. Wansing. Obligations, authorities, and history dependence. In H. Wansing, editor, *Essays on Non-classical Logic*, pages 247–258. World Scientific, 2001.

[24] Bernard Williams. Moral luck. In *Moral Luck*, pages 20–39. Cambridge University Press, 1982.