A logic of intention and attempt

Emiliano Lorini, Andreas Herzig

A Logic of Intention and Attempt

Emiliano Lorini (emiliano.lorini@istc.cnr.it)
Institut de Recherche en Informatique de Toulouse
118 route de Narbonne F-31062 Toulouse France

Andreas Herzig (herzig@irit.fr)
Institut de Recherche en Informatique de Toulouse
118 route de Narbonne F-31062 Toulouse France

Abstract. We present a modal logic called $\mathcal{LIA}$ (Logic of Intention and Attempt) in which we can reason about intention dynamics and intentional action execution. By exploiting the expressive power of $\mathcal{LIA}$, we provide a formal analysis of the relation between intention and action and highlight the pivotal role of attempt in action execution. Besides, we deal with the problems of *instrumental reasoning* and *intention persistence*.

Keywords: Intention, Action, Logic

1. Introduction

Since the seminal work of Cohen and Levesque (1990) aimed at implementing Bratman's philosophical theory of intention (Bratman, 1987), many formal logics for reasoning about mental attitudes of agents such as beliefs, desires and intentions have been developed. Among them we should mention the logics developed by Herzig and Longin, Konolige and Pollack, Meyer et al., Miller and Sandu, Rao and Georgeff, Shoham, Singh and Asher, Van Linder et al., Wooldridge (2004, 1993, 1999, 1997, 1991b, 1993, 1993, 1998, 2000). But, as far as we know, there are no formal logics in the literature which are able to combine in the same framework a precise description of both intention dynamics and intentional action execution. The problem of intentional action execution has been mainly investigated in the philosophical field where several authors have focused on the concepts of attempt, trying and volition. But, an adequate integration of these concepts into a general logical framework where we can reason about intention dynamics is still lacking. The main ambition of this work is to provide such a kind of integration by developing a sufficiently powerful multi-modal logic in which the following three fundamental problems for a theory of intentional action are addressed.

1. **Intentional action execution**: to have a precise characterization of the conditions under which an agent's intention transforms into an action, that is, to model the transition from a mental state called

intention to the performance of a movement, passing through a mental process which is called *trying* or *attempt*.

2. **Intention generation**: to individuate and to model appropriate principles of intention generation, that is, to model the mode of reasoning which is responsible for generating instrumental intentions and which is commonly called practical reasoning or instrumental reasoning.

3. **Intention persistence**: to account for persistence of intentions, that is, to have a precise characterization of the conditions under which an agent's intention persists over time.

While the third problem has received quite a lot of attention, the first and second problems have been rather neglected in the logical literature up to now. In this paper, we tackle all of them, going from the generation of an intention, to the performance of a bodily movement, via an attempt. The formal analysis developed in this work is aimed not only at clarifying some fundamental concepts in the theory of human intentional action such as the concepts of attempt and intention, but also at providing new insights for the design of artificial systems endowed with a body and a repertoire of movements that they can perform intentionally.

In the first part of the paper a general overview of the notion of attempt (or trying) is given (section 2). We focus on the particular form of attempt (or trying) conceived by some philosophers as synonymous of *volition*, where the term *volition* denotes the mental process which consists in an agent exerting his voluntary control over the initiation and the execution of a bodily movement. The relationships between this concept of volitional attempt and the classical notions of intention and basic action are discussed. In section 3, the syntax and the related semantics of the multi-modal logic $\mathcal{LIA}$ (Logic of Intention and Attempt) are presented. Special modal operators to talk about volitional attempts of agents are introduced in this section. These modal operators turn out to be crucial to explicitly represent the process going from an agent's proximal intention to the agent's performance. Moreover, modal operators to talk about beliefs and chosen goals of agents are given. The axiomatization of the logic is discussed in section 4. In section 5 a formal description of the relationship between attempt and mental attitudes is given and a solution to the formal characterization of the notion of action occurrence is proposed. In the second part of the paper we focus on dynamics of intentions (section 6). We begin with a formal analysis of the concepts of future-directed intention (distal intention), present-directed intention (proximal intention) and achievement goal.

Then, we discuss the problem of intention generation. We conclude with a discussion on the problem of persistence of intentions.

2. Conceptual foundations

2.1. Attempt as Volition

For some authors (Davis, 1979, Hornsby, 1980, O'Shaughnessy, 1973, Ginet, 1990) the words "trying" and "attempt" go proxy to what in the philosophical literature is called *volition*.[1] For instance, consider the Psycho-Psycho Law proposed by O'Shaughnessy (1973) in which the "bridging role" of *trying* between the mental level and the action execution level is explicitly stressed. He says:

> "...if an agent at an instant in time realizes that that instant is an instant at which he intends to perform action x, then logically necessarily he begins trying to do x at that very moment of realization..." (pp. 380).

Some authors have considered volition to be a mental action of an agent which is caused by an intention of the agent and which consists in the agent exerting his voluntary control over the initiation and the execution of a bodily movement (Davis, 1979, Ginet, 1990, Hornsby, 1980). In *An Essay concerning Human Understanding* Locke (1989) already provided an excellent analysis of this concept:

> "Volition, it is plain, is an act of the mind knowingly exerting that dominion it takes itself to have over any part of the man, by employing it in, or withholding it from, any particular action." (Book 2, XXI, 5; Book 2, XXI, 21).

In Hornsby's theory of action (Hornsby, 1980) *volition* (she calls it *trying*) is conceived as the most primitive actional notion. Her *trying* designates a mental action that is basic whereas all overt actions are nonbasic actions since they are always performed by trying to perform them. In Ginet's theory of Action (Ginet, 1990) a volition is not simply the trigger of a sequence of voluntary exertion of the body, but it is part of what it triggers.

While supposing that *volition* (*trying*) denotes a special kind of mental action which is responsible for causing a bodily movement, the previous authors distinguish it from mental states such as intentions and desires. A similar distinction is given by Searle (1983). Searle distinguishes proximal and distal intentions (*prior intentions*) from *intentions in action*, where the concept of intention in action has some strict similarities with the concepts of volition and trying of volitional theorists. At the heart of this distinction is the idea that only intentions

in action are *parts of* actions. On the one hand, effective prior intentions cause actions. On the other hand, effective intentions in action precede and cause bodily movements; and together with bodily movements that they cause, intentions in action constitute actions. In Searle's view complex motor patterns are represented in the prior intention. Once an agent intends to perform a complex bodily movement α and starts to perform it, simple bodily movements which make up the complex bodily movement α are the content of intentions in action and therefore they are performed intentionally. For example, a car driver may have the prior intention to change gears and while changing gears, he presses the clutch (the driver has the intention in action to press the clutch). According to Searle the driver is intentionally pressing the clutch even if he did not have the prior intention to do this.

Some authors (Proust, 2005, Pacherie, 2006) have developed cognitive models of what volitional theorists have commonly called trying, volition (or intention in action in the Searlian sense). Their aim is to explain how the initiating, sustaining and monitoring functions of the present-directed intention over its corresponding bodily movement become effective through a volition. These authors regard volition as a particular kind of *mental operation or procedure* (rather than a vague kind of mental action) which plays a precise functional role in a general architecture of the human mind. For instance, Proust (2005) has developed a teleological and functional model of volition. In her model a volition corresponds to a procedure of adaptive control in the sense of neuro-computational models of action control (Jordan and Wolpert, 1999). Once an agent conceives a proximal intention to perform a bodily movement α, this motivation gives rise to a process of adaptive control (a volition) aimed at performing α in a successful way. This process of adaptive control is a course of voluntary activity which consists both in the *selection* of appropriate motor programs for the successuful execution of the bodily movement α (motor planning) and in the *anticipation* of the effects of the selected motor programs (motor prediction), where such an anticipation provides an internal feedback for the correction of the ongoing bodily movement. For instance, suppose that the agent has a proximal intention to advance one step forward. This intention triggers a course of voluntary activity which consists in the selection of appropriate motor programs (e.g. moving the leg with a certain angle, direction, velocity; bending the knee in certain way; etc.), on the anticipation of the effects of the selected motor programs, and on the continuous adjustment of the body in order to advance one step forward in a successful way.

2.2. Intentional basic actions

The concept of basic action has been studied in philosophy by several authors (Danto, 1965, Goldman, 1970, Searle, 1983). Approximately, Searle (1983) states that a *basic action type* α of agent i is an action that i can intend to do without necessarily intending to do a different action β *in order to* do α. Therefore, if α is a *basic action type* of i then, i can intend to do α even if he lacks the beliefs about how he can do α (i.e. even if he does not have any cause-and-effect knowledge of the form "α may be done by doing β"). For instance, the action of lifting an arm, or the action of smiling are basic actions types of a normal agent i. In fact, i can intend to raise his arm, without intending to do a different action β in order to raise his arm. When i intends to raise his arm, he does not need to split the motor pattern "raise an arm" into more primitive constituents and to intend to do each of them. Agent i can intend to raise his arm, even if he does not have any cause-and-effect knowledge of the form "I can raise my arm by doing β". An action α successfully performed by i is a *basic action token* only if i intended to do α without intending to do a different action β in order to do α and α done by i is the instantiation of some basic action type of i. This implies that a *basic action token* is an action which is not performed by way of another action.
The basic actions types of an agent i could also be conceived as those motor patterns (i.e. bodily movements) which are in the repertoire of agent i (Davis, 1979). These are actions that agent i can do simply by intending to do them and without necessarily thinking how to do them. For example, in the repertoire of a normal agent i there are simple motor patterns such as "raising an arm", "moving a leg", but also complex motor patterns specialized for the accomplishment of specific tasks and the achievement of specific results such as "grasping an object", "tying one's shoes", "toggling the switch", etc. This perspective is also applicable to artificial settings, where an agent can be viewed as a system with a set of effectors related by a certain program. As in (Israel et al., 1991), the basic action types of the artificial agent would correspond to those bodily movements available to the effectors and the program, whereas basic action tokens of the agent would be movements *effected* by the artificial agent at a given moment. For instance, such an artificial agent could be either a robot with a real body (artificial limbs, rotating wheels, moving sensors, etc.) living in the real world or a robot with a simulated body living in a virtual environment.

Differently from basic actions, when an agent intends to do some *non-basic action* x, he necessarily intends to do a different action y *in order to* do x. Thus, as far as the mental aspect of non-basic actions is

concerned: if action x is non-basic for agent i, i can intend to do x only if he has a cause-and-effect knowledge of the way he can do x. As far as the executive aspect of a non-basic action is concerned: if an agent does a non-basic action x, he necessarily does x *by* doing a different action y, where the word "by" expresses a form of relation between actions which Goldman (1970) calls *generation*. This means that a non-basic action is an action that is performed by way of one or more actions.[2] There are non-basic actions which have a deep recursive structure. In fact, there could be a non-basic action x done by doing an action y which in turn is done by doing a further action z and so on. Such a decomposition of an agent's non-basic action x stops at the level of basic actions by their performance the agent does x. They therefore represent the agent's only direct intervention in the process of doing x. As Davidson puts it "the rest is up to nature" (Davidson, 1980). By way of example, consider Jack's non-basic action of killing Joe. Jack kills Joe by shooting him and Jacks shoots Joe by pulling the trigger of the gun. Jack's bodily movement of pulling the trigger (which consists in Jack's moving his forefinger in a certain way) is the only part of the non-basic action of killing Joe which is directly controlled by Jack. The effects of a non-basic action can be described by certain canonical forms *i brings it about that* φ and *after i brings it about that* ψ*, it is the case that* φ, where φ is the intrinsic result (Von Wright, 1963) of some non-basic action x, and ψ is the intrinsic result of some action y (either basic or non-basic) by doing which i does x. For example, Jack's non-basic action of killing Joe by shooting him can be described by the construction *Jack brings it about that Joe is shot* and *after Jack brings it about that Joe is shot, it is the case that Joe is dead*, where "Joe is dead" and "Joe is shot" are respectively the intrinsic result of Jack's action of killing Joe and the intrinsic result of Jack's action of shooting Joe.

2.3. The role of attempt in this work

The notions of attempt and trying formalized in the second part of this work (section 3) are the notions of volitional attempt and trying discussed in section 2.1. Only basic action types of an agent (i.e. bodily movements in an agent's repertoire) can be under the agent's voluntary control, that is, only basic action types of an agent can be the object of his volitions. Indeed, if α is the object of a volition of agent i (i.e. α is under the voluntary control of i), i does not need to think how to do α and to intend to something else in order to do α. Thus, by the definition of basic action given in section 2.2, it follows that α is the object of a volition of agent i only if α is a basic action type

of i. Non-basic actions such as killing someone, making a gift, etc. go beyond the voluntary control of agents and cannot be the object of their volitions. The basic action (i.e. bodily movement) by its performance an agent does a non-basic action x is the only part of x which is under the voluntary control of the agent. The rest of the non-basic action x goes beyond the agent's voluntary control. Therefore, the complete non-basic action x goes beyond the agent's voluntary control as well. For example, when Jack shoots Joe by performing the movement of pulling the trigger of the gun (which consists in Jack's moving his forefinger in a certain way), Jack only exerts voluntary control over the bodily movement of pulling the trigger. The complete non-basic action of shooting Joe goes beyond Jack's voluntary control. After Jack has moved his forefinger in a certain way, he just waits for the bullet to hit Joe without making any additional voluntary effort. These are the reasons why in this work we only deal with attempts of agents to do basic actions, that is, attempts of agents to perform bodily movements in their repertoire. We suppose that the expression "agent i tries to do α" means "agent i exerts his voluntary control over the performance of movement α", "i goes through the mental effort of moving his body in a certain way α".

We regard an agent i's *trying* (or *attempt*) *to do* α as the mental counterpart of a potentially performed bodily movement α in i's repertoire. We say *potentially* performed bodily movement since it is not always the case that if an agent tries to move his body in a certain way, he successfully moves his body in the way he tries. For example, while trying to raise his arm, an agent can encounter external obstacles which prevent him from raising his arm. We suppose that *i's attempt/trying to do* α denotes a mental process in agent i which consists in i exerting voluntary control over the initiation and the execution of the bodily movement α. In a way similar to volitional theorists, we suppose that *trying to do* α does not coincide with a proximal intention to do α. Indeed *trying to do* α is more than the disposition of intending to do α now. *Trying* already refers to the initiation of the basic action performance. More generally, an intention is a mental *state*, whilst an attempt is mental *process/action*. We accept the following statements as fundamental principles which relate basic actions and attempts.

PRINCIPLE 1. *If α is a basic action type of agent i and i has a proximal intention to do α, this intention proximally brings about i's attempt/trying to do α.*

PRINCIPLE 2. *If α is a basic action type of agent i and i tries/attempts to do α then such an attempt of i is brought about by i's proximal intention to do α.*

Principle 1 - that will be formalized in section 5 - is a refinement of O'Shaughnessy's Psycho-Psycho Law mentioned in section 2.1. It says that basic actions are actions which can be always *tried at will*, that is, a basic action type of i is a type of action whose occurrence can be controlled by the voluntary activity of i. To see this, imagine Jack intending to do the basic action of smiling so that other people will believe that he is happy. In this scenario, Jack's intention brings about Jack's attempt to smile which consists in Jack exerting his voluntary control over the corresponding movement of the mouth. It has to be noted that, Principle 1 prevents from identifying as a basic action type of an agent some movement or change in his body that the agent can produce in a "fortuitous" way but which is not under his voluntary control. Imagine a man conceiving an intention to speed up his heartbeat. As a consequence, he becomes so excited that his heartbeat speeds up. Being something that the man does in a fortuitous way, we cannot consider "speeding up the heartbeat" as a basic action type of the agent. Indeed, it is not always the case that if he intends to speed up his heartbeat then this intention necessarily triggers appropriate actional mechanisms specialized for speeding up the heartbeat and a course of voluntary control over this event. Under normal conditions, even if the man intends to speed up his heartbeat here and now, he is not able to exert a voluntary control over the movements of his heart. Therefore, he cannot try to speed up his heartbeat.

According to Principle 2, given a basic action type α of agent i (i.e. a bodily movement in i's repertoire), i's attempt to perform α is necessarily caused by i's proximal intention to peform α. In fact, an agent's non-intentional behavior such as a reactive response to a given stimulus does not involve any attempt. When agent i performs some movement in a spontaneous and automic way, he does not really try to perform that movement (i.e. i does not exert any voluntary control over the performance of that movement). For instance, imagine agent i hearing a sudden pistol shot behind him, and quickly turning the head. In this scenario i's behavior does not involve any volitional attempt to turn the head. In fact, i does not exert any voluntary control over the movement of turning the head, he does not go through any mental effort. He simply turns the head in a fast and automatic way.

The present work is also devoted to investigate the formal relationships between attempt and intention. Here, we briefly summarize the relevant categories of intention which will intervene in the second part of the paper (section 3) where the logic of intention and attempt $\mathcal{LIA}$ will be presented. In a way similar to Bratman (1987) and Mele (1992) we specify two categories inside the general class of intentions: *future-directed intentions* (or *distal intentions*) and *present-directed intentions*

(or *proximal intentions*).[3] A *future-directed intention* is the intention to do some basic action later whereas a *present-directed intention* is an intention to do some basic action here and now. In a way similar to Bratman, we suppose that a future-directed intention is an element of a partial plan, that is the input of practical reasoning aimed at filling or modifying this partial plan. Present-directed intentions are direct motivations to do the basic action when the time point of the planned action execution is attained. Finally, we have a mental process called *attempt to do* α (or *trying to do* α) which, as emphasized above, is always caused by a present-directed intention to perform α and consists in an agent exerting voluntary control over the initiation and the execution of the bodily movement α.[4]

2.4. Other conceptions of attempt and trying

As argued in section 2.3, the concept of volitional attempt and trying only concern basis action types of agent corresponding to bodily movements in the agent's repertoire. This concept does not apply to non-basic actions such as killing someone, illuminating a room, making a gift, etc. which go beyond the voluntary control of an agent. But this is not the only way *trying* and *attempt* have been defined. There are alternative conceptions of these notions in literature which are not taken into account in this paper. Under these alternative conceptions it makes sense to say that "an agent tries to kill someone", "an agent tries to illuminate the room", etc. According to Sellars (1967) for instance, tryings are not volitions and volitions are not tryings. In his view *trying* corresponds to a particular mental configuration of beliefs about a possible failure which occurs when an action is initiated. Approximately, Sellars holds that *i tries to do* α if and only if *i* does one or more things without being sure that they will grow into a doing α. According to Schroeder (2001), the applicability of the word *trying* depends above all one the external observer's epistemic stance: whether he regards the action of a certain agent as a possible failure. When an external observer speaks of someone *trying to do* α, then he must leave room for a possible failure of the action execution. Differently from Sellars, in Schroeder's conception of *trying* the expectation of a possible failure is an expectation of the external observer rather than an expectation of the performing agent. This shows that Schroeder has in mind a concept of "not necessarily successful action". In order to prevent misunderstandings, we call *trying*/attempt** the kind of trying and attempt studied by Sellars and *trying**/attempt*** the kind studied by Schroeder. With these notions of *trying** and *trying***, it is possible to specify the meaning of the expressions "agent *i tries** to do

a non-basic action α" and "agent i *tries*** to do a non-basic action α". For instance, when saying that Jack *tries** to kill Joe, we are saying that Jack does some action now (e.g. Jack shoots Joe) and Jack is not sure that after he performs such an action (e.g. after he shoots Joe), it is the case that Joe is dead. When saying that Jack *tries*** to kill Joe, we are saying that Jack does some action now (e.g. Jack shoots Joe) and it is possible that after Jack does such an action (e.g. after Jack shoots Joe), it is not the case that Joe is dead.

2.5. Related works: attempt and trying in logic

As far as we know, there are few logical theories of action and intention in the literature which are able to characterize the conditions under which an agent's intention transforms into a bodily movement and to model the transition from an intention to the performance of a movement, passing through a mental process which is called *trying* or *attempt*. Most of logics of action and intention (see for instance Cohen and Levesque, Herzig and Longin, Meyer et al., Wooldridge, 1990, 2004, 1999, 2000) generalize the operator $[\alpha]$ of propositional dynamic logic (PDL) (Harel et al., 2000) - which was introduced for describing the effects of executing a certain program α - to operators of the form $[i : \alpha]$ describing the effects of an action α peformed by a certain agent i.[5]. Even if they are able to characterize the notion of action occurrence, all these logics of action and intention lack a fine-grained notion of volitional attempt. Nevertheless, there are some logicians working on the formalization of intentional action who have been attracted by this notion. For instance, Rao and Georgeff (1991b) have developed a logic of intention and branching time where two modal operators $succeeded(\alpha)$ and $failed(\alpha)$ are used. The former reads "the agent has failed in doing action α" whilst the latter reads "the agent has succeeded in doing action α". From these two operators, Rao and Georgeff introduce the abbreviation $done(\alpha) =_{def} succeeded(\alpha) \vee failed(\alpha)$ which allows them to formalize something similar to a concept of volitional attempt. In their logic, $done(\alpha)$ expresses that "the agent has either failed or succeeded in doing action α", that is, "he has tried to do action α". There is a difference between the formalization of *attempt* presented in this paper and Rao and Georgeff's. In our logic *attempt* is a primitive notion which is defined by a special modal operator for *attempt*. Moreover, in our approach *action success* and *action failure* can be defined from the primitive notion of *attempt*. In Rao and Georgeff's approach *attempt* can be defined on the basis of the concepts (and related two modal operators) of *action failure* and *action success*. In section 5.2 we will show that the two

approaches are equivalent in expressivity. Nevertheless, we think that *Occam's razor* is the main reason for preferring a logic with a single modal operator for *attempt* such as the one presented in this paper to a logic with a modal operator for *action success* and a modal operator for *action failure* such as the one proposed by Rao and Georgeff.

Santos et al., Santos et al. (1997a, 1997b) have appropriately extended the logic of agency developed in (Kanger, 1971, Pörn, 1977) and refined in (Elgesem, 1993) by introducing an operator for representing not necessarily successful actions. Santos et al. use the modal operator E_i, where expressions of the form $E_i\varphi$ are read "agent i brings it about that φ". In addition to E_i, they use the modal operator H_i, where expressions of the form $H_i\varphi$ are read "agent i attempts to make it the case that φ". The main difference between E_i and H_i is that $E_i\varphi \rightarrow \varphi$ is valid, while $H_i\varphi \rightarrow \varphi$ is not. Santos et al.'s notion of trying is similar to Schroeder's notion discussed in section 2.4. As stressed in section 2.4, this is radically different from the concept of volitional attempt we are aimed at investigating in this paper. Indeed, our notion of volitional attempt is a mentalistic notion, that is, it denotes some action/process occurring in the mind of an agent. The notion of attempt as "not necessarily successful action" studied by Santos et al. does not have this status. It only describes something which makes sense in the perspective of an external observer who looks at an agent engaged in an action performance and prospects the possibility that the agent will fail to do such an action.

3. Formal logic

3.1. General overview

The rest of the paper is devoted to develop a formal logic in which the concepts presented in the previous sections are formalized and their properties investigated. We call our logic $\mathcal{LIA}$: Logic of Intention and Attempt. In $\mathcal{LIA}$ the basic constructs of propositional dynamic logic (PDL) (Harel et al., 2000) are reinterpreted in order to formalize some fundamental aspects of the general theory of intentional action such as the notion of volitional attempt which have been neglected in the logical literature on action and intention up to now. In particular, an *atomic action* (or *atomic program*) in the sense of PDL is conceived in $\mathcal{LIA}$ as a basic action type of one or more agents, that is, as a kind of bodily movement in the actional repertoire of one or more agents. Basic actions types are denoted in $\mathcal{LIA}$ by symbols $\alpha, \beta, ...$ Moreover, the standard PDL operators of the form $[\alpha]$ are substituted in $\mathcal{LIA}$

with attempt operators of the form $[\![i : \alpha]\!]$ (whose duals are $\langle\langle i : \alpha \rangle\rangle$) associating agents with basic action types. The $\mathcal{LIA}$ operator $[\![i : \alpha]\!]$ is supposed to describe the effects of agent i's attempt to perform movement α. As we have stressed in section 2.3, an *attempt/trying to do* α *of agent* i is regarded here as synonymous of volition, i.e. a mental procedure which consists in agent i exerting his voluntary control over the initiation and the execution of the bodily movement α.

As emphasized in section 1, the issue of intentional action execution (i.e. how an agent's intention transforms into an action) has been rather neglected in the logical literature up to now. We think that the concept of attempt formalized in this paper is crucial for the clarification of such an intriguing issue. We also think that $\mathcal{LIA}$ can provide new insights into the field of logical modelling of artificial settings where an agent is conceived as a robot with a real body (artificial limbs, rotating wheels, moving sensors, etc.) interacting with the real world or a robot with a simulated body interacting with a virtual environment. The basic action types of the robot would correspond to those bodily movements available to the robot's effectors and the program controlling the robot's behavior. The robot attempt's to perform a certain movement α would be the command of executing movement α to the robot's effectors by the program.

In this work we will not consider non-basic actions such as killing someone or making a gift which, as pointed out in section 2.2, go beyond the voluntary control of an agent and can be described by certain canonical forms *i brings it about that* φ and *after i brings it about that* ψ*, it is the case that* φ. There are several logics in the literature in which the notion of non-basic action can be formalized. These are the so-called "bringing it about" (or "seeing to it that") logics of action and agency which are somehow complementary to PDL. In fact, differently from PDL, in these languages there are no symbols $\alpha, \beta, ...$ standing for names of actions, and what an agent does is only described in terms of the result that the agent brings about by his acting. For instance in the formal framework developed by Kanger, Pörn (1971, 1977) and refined by Elgesem (1993), an operator E_i is introduced. In STIT theory (Belnap et al., 2001, Horty and Belnap, 1995) the modal operator $[i\ dstit :]$ is given, called deliberative STIT operator. Even if different in their semantics, the operators E_i and $[i\ dstit :]$ are aimed at capturing the same property. Namely, they are used to express the fact that an agent ensures a result by acting in a certain way. In Kanger's logic the formula $E_i\varphi$ is read "agent i brings it about that φ", whereas in STIT theory a formula $[i\ dstit : \varphi]$ is read "i sees to it that φ". In Segerberg's logic of *bringing it about* (Segerberg, 1989) an operator δ is introduced, where $\delta\varphi$ denotes actions leading to states where φ holds and a formula

$[\delta\varphi]\,\psi$ is read "after an agent brings it about that φ, it is the case that ψ". In Segerberg's logic the recursive structure of a non-basic action can be easily captured. For example, Jack's action of killing Joe by shooting him can be described by the formula $[\delta JoeShot]\,JoeDead$. The main problem with logics of "bringing it about" (and "seeing to it that") is that they do not provide a clear distinction between an agent's basic actions (*alias* bodily movements in the actional repertoire of the agent) and his non-basic actions (*alias* actions which go beyond his voluntary control). In particular, they do not have an explicit characterization of bodily movements that an agent brings about by exerting voluntary control over their performance. This is the reason why in this work we prefer to adopt a perspective on the problem of action which is similar in spirit to PDL, and to include in the language a set of action labels $\alpha, \beta, ...$ in order to have an explicit reference to basic action types of agents and to be able to model volitional attempts.

In $\mathcal{LIA}$ a formula $[\![i : \alpha]\!]\,\varphi$ stands for "after agent i tries to move his body in a certain way α, it is the case that φ holds", that is, "after i goes through the mental effort of moving his body in a certain way α, it is the case that φ holds". For example, $[\![Bill : raiseArm]\!]\,BillArmUp$ stands for "after Bill tries to raise his arm, it is the case that Bill's arm is up", that is, "after Bill goes through the mental effort of raising his arm, it is the case that Bill's arm is up". Similarly, a formula $\langle\!\langle i : \alpha \rangle\!\rangle \varphi$ stands for "agent i tries to move his body in a certain way α and φ holds after i's attempt", that is, "i goes through the mental effort of moving his body in a certain way α and φ holds after i's attempt". Thus, when the two formulas $\langle\!\langle i : \alpha \rangle\!\rangle \top$ and $\langle\!\langle j : \alpha \rangle\!\rangle \top$ are true it does not mean that the two agents i and j try to bring about the same result α. It means that agent i tries to move his body in a certain way α and j tries to move his body in the same way α. Since i and j are different agents (with different bodies and at different places), i's attempt to perform movement α and j's attempt to perform the same movement α are supposed to be distinct events (on this point see also sections 5.2.1 and 6.4).

$\mathcal{LIA}$ has also the *henceforth* modal operator $\Box$ of standard temporal logic and modal operators for mental attitudes of the form Bel_i and $Goal_i$. The former are standard doxastic operators and express what agents currently believe. The latter refer to chosen goals of agents, i.e. goals that agents decide to pursue. It is supposed that an agent cannot have conflicting chosen goals (i.e. he cannot have the chosen goal that φ and the chosen goal that $\neg\varphi$) and that an agent's chosen goals must be consistent with his beliefs.

3.2. Syntax

Let $\Pi = \{p, q, ...\}$ be a set of atomic formulas, $AGT = \{i, j, ...\}$ a non-empty finite set of agents, $ACT = \{\alpha, \beta, ...\}$ a non-empty finite set of names of bodily movements. Elements in ACT are supposed to denote simple and complex motor patterns (i.e. bodily movements) such as "lifting an arm", "moving a sensor", "grasping an object", etc. For the sake of simplicity, we suppose here that every agent in AGT has ACT as repertoire of bodily movements, that is, every $\alpha \in ACT$ is a basic action type of every agent $i \in AGT$. Under this assumption, the language $\mathcal{L}_{\mathcal{LIA}}$ includes all formulas $\langle\langle i : \alpha \rangle\rangle \varphi$ where $\alpha \in ACT$ and $i \in AGT$, that is, $\mathcal{L}_{\mathcal{LIA}}$ is defined as the smallest superset of Π such that:

- if $\varphi, \psi \in \mathcal{L}_{\mathcal{LIA}}$ and $i \in AGT$ then $\neg\varphi, \varphi \vee \psi, \Box\varphi, Bel_i\varphi, Goal_i\varphi \in \mathcal{L}_{\mathcal{LIA}}$;
- if $\alpha \in AGT$, $i \in AGT$ and $\varphi \in \mathcal{L}_{\mathcal{LIA}}$ then $\langle\langle i : \alpha \rangle\rangle \varphi \in \mathcal{L}_{\mathcal{LIA}}$.[6]

$Bel_i\varphi$ is read "agent i believes that φ" whereas $Goal_i\varphi$ is read "agent i has the chosen goal that φ" or simply "agent i wants that φ". $\Box\varphi$ is read "φ is true in the present and will always be true". Since future is linear in our logic attempts are conceived as transitions from one time point to the future next. As emphasized in section 3.1, $\langle\langle i : \alpha \rangle\rangle \varphi$ should be read "agent i tries to move his body in a certain way α and φ holds after i's attempt". For the sake of simplicity, we will shorten this to "agent i tries to do α and φ holds after i's attempt". Hence $\langle\langle i : \alpha \rangle\rangle \top$ has to be read "i tries to do α" or "i attempts to do α". Several abbreviations are used in our logic. The classical Boolean connectives $\wedge$, $\rightarrow$, $\leftrightarrow$, $\top$ (tautology) and $\bot$ (contradiction) are defined from $\vee$ and $\neg$ in the usual manner. Moreover, $[\![i : \alpha]\!] \varphi$ abbreviates $\neg\langle\langle i : \alpha \rangle\rangle\neg\varphi$, $\Diamond\varphi$ abbreviates $\neg\Box\neg\varphi$. $[\![i : \alpha]\!] \varphi$ has to be read "after i tries to do α, it is the case that φ" or "φ holds after i tries to do α". Hence $[\![i : \alpha]\!] \bot$ has to be read "agent i does not try to do α". Finally, $\Diamond\varphi$ has to be read "φ will eventually be true".

3.3. Basic semantics

The class of models $\mathbf{M}$ for $\mathcal{L}_{\mathcal{LIA}}$ is the set of 6-tuples of the form $M = (W, R_\Box, B, G, R, V)$ where:

- W is a set of possible worlds or states;
- $R_\Box$ is a binary relation on W;
- B is a collection of binary relations B_i on W one for every $i \in AGT$;

- G is a collection of binary relations G_i on W one for every $i \in AGT$;
- R is a collection of binary relations $R^{att}_{i:\alpha}$ on W one for every couple $i:\alpha$ where $i \in AGT$ and $\alpha \in ACT$;
- $V : \Pi \longrightarrow 2^W$ is a valuation function.

Informally, given an arbitrary world w, $R_\Box(w)$ is the set of worlds that are in the future of w, $B_i(w)$ is the set of worlds that i considers possible at w, $G_i(w)$ is the set of worlds which are compatible with i's chosen goals at w, $R^{att}_{i:\alpha}(w)$ is the set of worlds which are accessible from world w via i's attempt to do α.

Given a model M, a world w and a formula φ, we write $M, w \vDash \varphi$ to mean that φ is true at world w in M, under the basic semantics. Truth conditions for propositional atoms, negation and disjunction are entirely standard. The following are truth conditions for $Bel_i\varphi$, $Goal_i\varphi$, $\langle\langle i:\alpha\rangle\rangle\varphi$ and $\Box\varphi$.

- $M, w \vDash Bel_i\varphi \Longleftrightarrow \forall w'$ if $w' \in B_i(w)$ then $M, w' \vDash \varphi$.
- $M, w \vDash Goal_i\varphi \Longleftrightarrow \forall w'$ if $w' \in G_i(w)$ then $M, w' \vDash \varphi$.
- $M, w \vDash \Box\varphi \Longleftrightarrow \forall w'$ if $w' \in R_\Box(w)$ then $M, w' \vDash \varphi$.
- $M, w \vDash \langle\langle i:\alpha\rangle\rangle\varphi \Longleftrightarrow \exists w'$ such that $w' \in R^{att}_{i:\alpha}(w)$ and $M, w' \vDash \varphi$.

In the next section we restrict the set of models by constraints on the accessibility relations, and give an axiomatization of the resulting set of models.

4. Axiomatization

We give the following axiom and inference rule schemas.

- 0. All tautologies of propositional calculus.
- **Axioms of Time and Attempt**.
  1a. $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$
  2a. $\Box\varphi \to \varphi$
  3a. $\Box\varphi \to \Box\Box\varphi$
  4a. $\Diamond\varphi \wedge \Diamond\psi \to \Diamond(\varphi \wedge \Diamond\psi) \vee \Diamond(\psi \wedge \Diamond\varphi)$
  5a. $[\![i:\alpha]\!](\varphi \to \psi) \to ([\![i:\alpha]\!]\varphi \to [\![i:\alpha]\!]\psi)$
  6a. $\langle\langle i:\alpha\rangle\rangle\varphi \to [\![j:\beta]\!]\varphi$
  7a. $\Box\varphi \to [\![i:\alpha]\!]\varphi$
  8a. $\varphi \wedge \langle\langle i:\alpha\rangle\rangle\Box\varphi \to \Box\varphi$

– **Axioms of Mental Attitudes**.
1b. $Bel_i(\varphi \rightarrow \psi) \rightarrow (Bel_i\varphi \rightarrow Bel_i\psi)$
2b. $Bel_i\varphi \rightarrow Bel_iBel_i\varphi$
3b. $\neg Bel_i\varphi \rightarrow Bel_i\neg Bel_i\varphi$
4b. $Goal_i(\varphi \rightarrow \psi) \rightarrow (Goal_i\varphi \rightarrow Goal_i\psi)$
5b. $Goal_i\varphi \rightarrow Bel_iGoal_i\varphi$
6b. $\neg Goal_i\varphi \rightarrow Bel_i\neg Goal_i\varphi$
7b. $Goal_i\varphi \rightarrow \neg Bel_i\neg\varphi$

– **Axioms of mental attitude dynamics**.
1c. $Bel_i [\![j:\beta]\!] \varphi \rightarrow [\![j:\beta]\!] Bel_i\varphi \vee Bel_i [\![j:\beta]\!] \bot$
2c. $[\![j:\beta]\!] Bel_i\varphi \rightarrow Bel_i [\![j:\beta]\!] \varphi \vee [\![j:\beta]\!] \bot$
3c. $Bel_i(\Box Bel_i\varphi \rightarrow Bel_i\Box\varphi)$
4c. $Goal_i [\![j:\alpha]\!] \varphi \wedge [\![j:\alpha]\!] Bel_i\varphi \rightarrow [\![j:\alpha]\!] Goal_i\varphi \vee Goal_i [\![j:\alpha]\!] \bot$
5c. $\langle\!\langle i:\alpha \rangle\!\rangle \top \rightarrow Goal_i \langle\!\langle i:\alpha \rangle\!\rangle \top$
6c. $[\![i:\alpha]\!] \bot \rightarrow Goal_i [\![i:\alpha]\!] \bot$

– **Inference rules**
R1. $\frac{\vdash\varphi \; \vdash\varphi\rightarrow\psi}{\vdash\psi}$ (Modus Ponens)
R2. $\frac{\vdash\varphi}{\vdash\Box\varphi}$ ($\Box$-Necessitation)
R3. $\frac{\vdash\varphi}{\vdash[\![i:\alpha]\!]\varphi}$ ($[\![i:\alpha]\!]$-Necessitation)
R4. $\frac{\vdash\varphi}{\vdash Bel_i\varphi}$ (Bel_i-Necessitation)
R5. $\frac{\vdash\varphi}{\vdash Goal_i\varphi}$ ($Goal_i$-Necessitation)

The rest of the section contains explanations of the axioms together with semantic constraints guaranteeing their validity, and a completeness theorem.

4.1. Properties of time and attempt

Axioms 1a-4a and rule of inference R2 define a S4.3 logic for the temporal operator $\Box$ (Goldblatt, 1992). Axiom 5a and rule of inference R3 define a minimal normal modal logic for attempt operators of the form $[\![i:\alpha]\!]$. They thus do not have an associated semantic constraint. Axioms 2a and 3a express the interpretation of $R_\Box$ by a reflexive and transitive relation, that is, for every $w \in W$:

2A. *Reflexivity* of $R_\Box$: $w \in R_\Box(w)$

3A. *Transitivity* of $R_\Box$: if $w' \in R_\Box(w)$ and $v \in R_\Box(w')$ then $v \in R_\Box(w)$.

Axiom 4a corresponds to the following semantic constraint. For every $w \in W$:

4A. if $w' \in R_\Box(w)$ and $w'' \in R_\Box(w)$ then $w'' \in R_\Box(w')$ or $w' \in R_\Box(w'')$.

It follows from this last condition that time is not branching towards the future. Axiom 6a ensures that all attempts occurring in a state w correspond to transitions to the same state v, that is, for any $\alpha, \beta \in ACT$, $i, j \in AGT$ and $w \in W$, it holds that:

6A. if $w' \in R^{att}_{i:\alpha}(w)$ and $w'' \in R^{att}_{j:\beta}(w)$ then $w' = w''$.

From the previous semantic constraint it follows that attempts are deterministic, that is, for any $\alpha \in ACT$, $i \in AGT$ and $w \in W$:

– if $w' \in R^{att}_{i:\alpha}(w)$ and $w'' \in R^{att}_{i:\alpha}(w)$ then $w' = w''$.

Axiom 7a connects time and attempt, and corresponds to the following constraint. For any $\alpha \in ACT$, $i \in AGT$ and $w \in W$:

7A. $R^{att}_{i:\alpha}(w) \subseteq R_{\Box}(w)$.

Thus worlds resulting from an attempt to do α in w are in the future of w. Finally, Axiom 8a ensures that an attempt cannot "jump" to a distant future world that is more than one time step away, i.e. if a world w' is accessible from world w via an attempt to do α then every future world w'' different from w is in the future of w'. Formally, for any $\alpha \in ACT$, $i \in AGT$ and $w \in W$:

8A. if $w' \in R^{att}_{i:\alpha}(w)$ and $w'' \in R_{\Box}(w)$ and $w \neq w''$ then $w'' \in R_{\Box}(w')$.

This constraint ensures that there is no third future world between a world w and the outcome w' of an attempt starting at w.

4.1.1. *Discussion*

The formal properties of models given in this section characterize a semantics that is weaker than that of linear temporal logic LTL. Differently from standard linear temporal logic (Gabbay et al., 1980), we do not use a relation $R_{\bigcirc}$ for interpreting a modal *next* operator $\bigcirc$ - where $R_{\bigcirc}(w)$ is the immediate successor of w-, and for building the relation $R_{\Box}$ as the reflexive and transitive closure of $R_{\bigcirc}$. In $\mathcal{LIA}$ the relation for the *next* operator is replaced by a set of relations for attempts. Every attempt corresponds to a time step, and several attempts can occur in parallel. Moreover, we have supposed that all attempts of the same agent and all attempts of different agents occurring in a world w lead to the same world. This implies that all attempts of the same agent and all attempts of different agents starting in a world w occur in parallel. This explains why we have phrased $\langle\langle i : \alpha \rangle\rangle \varphi$ "agent i tries to do α and φ holds after this attempt" rather than "*it is possible that* agent i tries to do α and φ holds after this attempt". Thus, in our semantics, when an agent i tries to do two different actions α and β

at a world w, all effects of i's attempt to do α and all effects of i's attempt to do β are effects of the joint occurrence of the two attempts of i; when i and j are different agents, i tries to do α and j tries to do β, all effects of i's attempt to do α and all effects j's attempt to do β are effects of the joint occurrence of the two attempts of i and j. For instance, suppose that an agent called $Bill$ is at world w and tries to perform the coordinated hand movement of grasping and he succeeds in performing this movement. There is an object 1 ($o1$) in front of Bill in such a way that object 1 will be in Bill's hand after he tries to grasp at w: $M, w \models \langle\langle Bill : grasp\rangle\rangle o1InBillHand$. Suppose that there is a second agent called Bob who also tries to grasp at w and succeeds in performing this movement. A certain object 2 ($o2$) is in front of Bob in such a way that object 2 will be in Bob's hand after he tries to grasp at w: $M, w \models \langle\langle Bob : grasp\rangle\rangle o2InBobHand$. Given that in our logic attempts of different agents correspond to transitions to the same world, we have that $M, w \models \langle\langle Bill : grasp\rangle\rangle(o1InBillHand \wedge o2InBobHand)$ and $M, w \models \langle\langle Bob : grasp\rangle\rangle(o1InBillHand \wedge o2InBobHand)$. This means that the conjunction $o1InBillHand \wedge o2InBobHand$ is the effect of the joint occurrence of Bill's attempt to grasp and Bob's attempt to grasp. This assumption about *deterministic* effects of attempts comes with the assumption of linearity of time. Indeed, in the present analysis, we suppose that time evolves linearly without branching. Since attempts occur over time, they must necessarily follow the unique temporal line. It has to be noted that this assumption would not be acceptable when attempting and trying are conceived as synonymous of "not necessarily successful action" (see sections 2.4 and 2.5 for a discussion on this alternative conception of attempt). Indeed, a "not necessarily successful action" is an action which can either turn out to succeed or turn out to fail. So, by definition, it is an action whose outcomes are non-deterministic. Again we want to stress that in this paper we focus on a different concept of attempt. Our aim is to provide a formal characterization of the concept of volitional attempt discussed in sections 2.1 and 2.3. We think that such a concept can be properly formalized in a logic in which linear time is assumed and in which attempts of the same agent and attempts of different agents starting from the same world w occur in parallel and are transitions to the same world w'.

4.2. Static properties of mental attitudes

We adopt the system KD45 for modelling beliefs and the system KD for modelling chosen goals. Thus, we accept omniscience for both kinds of mental attitudes, i.e. an agent's beliefs and chosen goals are closed

under tautologies, conjunction, and logical consequences. Axioms 1b and 4b with rules of inference R4 and R5 correspond to a minimal normal modal logic for modal operators Bel_i and $Goal_i$. Axioms 2b and 3b express the interpretations of every B_i by transitive and euclidean relations, that is, for any $i \in AGT$ and $w \in W$:

2B. *Transitivity* of B_i: if $w' \in B_i(w)$ and $v \in B_i(w')$ then $v \in B_i(w)$

3B. *Euclideanity* of B_i: if $v, v' \in B_i(w)$ then $v' \in B_i(v)$ and $v \in B_i(v')$

According to Axiom 7b chosen goals of an agent must be consistent with his beliefs. Such an axiom corresponds to a *weak realism* principle requiring that there is always a world which is both compatible with agent i's beliefs and with agent i's chosen goals, that is, for any $i \in AGT$ and $w \in W$ it holds that:

7B. $B_i(w) \cap G_i(w) \neq \emptyset$.

Stronger principles have been proposed in the logical literature on action and intention. For instance, in Cohen and Levesque, Herzig and Longin, Miller and Sandu (1990, 2004, 1997) the principle $Bel_i\psi \rightarrow Goal_i\psi$ is adopted (i.e. if an agent believes that ψ holds then, he wants that ψ holds), whereas Shoham (1993) adopts the principle $Goal_i\psi \rightarrow Bel_i\psi$ (i.e. if an agent wants that ψ holds then, he believes that ψ holds). These two principles have been criticized by Rao and Georgeff, Wooldridge, Bratman (1991a, 2000, 1987). By choosing Axiom 7b instead of the stronger $Bel_i\psi \rightarrow Goal_i\psi$ and $Goal_i\psi \rightarrow Bel_i\psi$, we impose a minimal constraint which makes us able to adhere to psychological realism. Indeed, a rational agent cannot decide to pursue a goal if he believes that this goal cannot be achieved. According to the semantic property 7B corresponding to Axiom 7b, the intersection between $B_i(w)$ and $G_i(w)$ is never empty. It follows that both B_i and G_i are serial relations. Formally, for any $i \in AGT$ and $w \in W$: $B_i(w) \neq \emptyset$ and $G_i(w) \neq \emptyset$. Therefore, an agent cannot have contradictory beliefs nor conflicting goals (i.e. $\vdash_{\mathcal{LIA}} \neg(Bel_i\varphi \wedge Bel_i\neg\varphi)$ and $\vdash_{\mathcal{LIA}} \neg(Goal_i\varphi \wedge Goal_i\neg\varphi)$ are derivable in our logic). Axiom 5b and 6b are axioms of positive introspection and negative introspections for chosen goals. These two principles are assumed by Dunin-Keplicz and Verbrugge (2002) as well. They correspond to the following two semantic constraints. For any $i \in AGT$ and $w \in W$ we have that:

5B. if $w' \in B_i(w)$ then $G_i(w') \subseteq G_i(w)$.

6B. if $w' \in B_i(w)$ then $G_i(w) \subseteq G_i(w')$.

4.3. Dynamics of beliefs

Belief dynamics are modelled in our logic by means of Axioms 1c and 2c. They are respectively the axiom of *no forgetting* (NF axiom) and the axiom of *no learning* (NL axiom) for beliefs. Sometimes they have also been called *perfect recall* and *no miracles* (Van Benthem and Pacuit, 2006). A lot of researchers have studied similar principles for the interaction between belief and action or between knowledge and action. Among them we should mention Baltag et al., Fagin et al., Gerbrandy, Scherl and Levesque (1998, 1995, 1999, 2003). Axioms 1c and 2c correspond to the following two semantic properties. For any $i, j \in AGT$, $\alpha \in AGT$ and $w \in W$ it holds that:

1C. if $(B_i \circ R^{att}_{j:\alpha})(w) \neq \emptyset$ then $(R^{att}_{j:\alpha} \circ B_i)(w) \subseteq (B_i \circ R^{att}_{j:\alpha})(w)$;

2C. if $R^{att}_{j:\alpha}(w) \neq \emptyset$ then $(B_i \circ R^{att}_{j:\alpha})(w) \subseteq (R^{att}_{j:\alpha} \circ B_i)(w)$;

where $\circ$ is the standard composition operator between two binary relations: $(R^{att}_{j:\alpha} \circ B_i)(w) = \bigcup\{B_i(b) : v \in R^{att}_{j:\alpha}(w)\}$. Given the determinism of actions (constraint 6A), 1C and 2C are equivalent together to:

– if $w' \in R^{att}_{j:\alpha}(w)$ and $(B_i \circ R^{att}_{j:\alpha})(w) \neq \emptyset$ then $B_i(w') = (B_i \circ R^{att}_{j:\alpha})(w)$.

In accepting the NF and NL axioms for beliefs, we suppose that events are always uninformative, that is, apart from the mere occurrence of j's attempt to do α at w, i should learn nothing about the particular effects of j's attempt to do α that obtain in w'. What an agent i believes at a world w' only depends on what the agent believed at the previous world w and on the attempt which has occurred and which is responsible for the transition from w to w'. Besides, the two axioms rely on an additional assumption of complete and correct information. It is supposed that j's attempt to do α occurs if and only if every agent is informed of this fact. Hence all occurrences of attempts are public. Finally, it has to be noted that the previous two axioms do not say anything about belief revision. They only describe belief dynamics in conditions of no surprise, that is, when the occurring attempt is not unexpected by an agent. In fact, the two axioms do not specify how agent i's beliefs evolve from w to w' when j's attempt to do α is responsible for the transition from w to w' (i.e. $w' \in R^{att}_{j:\alpha}(w)$) and, at w agent i expected j's attempt to do α not to occur (i.e. $(B_i \circ R^{att}_{j:\alpha})(w) = \emptyset$).

A further axiom concerning beliefs is 3c. It is a subjective version of the previous NL axiom for beliefs. Its semantic counterpart is the following. For any $i \in AGT$ and $w \in W$:

3C. if $v \in B_i(w)$ then $(R_\Box \circ B_i)(v) \subseteq (B_i \circ R_\Box)(v)$.

According to property 3C, if v is a world that agent i considers possible at world w then, if w' is a future world of a world w'' that agent i considers possible at world v then there is a world v' which is a future world of v such that w' is a world that agent i considers possible at world v'.

4.4. Dynamics of chosen goals

According to Axiom 4c, if i wants φ to be true after j tries to do α, i does not want j not to try to do α and, if after j tries to do α i will believe that φ then, after agent j tries to do α, i will want φ to be true. The semantic property corresponding to this axiom can be easily specified. Suppose the actual world is w, and agent j tries to do α leading to a new actual world $w' \in R^{att}_{j:\alpha}(w)$. Now, consider an arbitrary world w'' which is compatible with i's chosen goals at world w': $w'' \in G_i(w')$. Then, either i makes mentally happen j's attempt to do α in one of his preferred worlds at w and finally collects the resulting world w'': $w'' \in (G_i \circ R^{att}_{j:\alpha})(w)$, or w'' is a world which is compatible with i's beliefs at w': $w'' \in B_i(w')$. We thus have $G_i(w') \subseteq (B_i \circ R^{att}_{j:\alpha})(w) \cup (R^{att}_{j:\alpha} \circ G_i)(w)$. We restrict this relation in order to avoid that the set of i's preferred worlds at w' is necessarily a subset of the set of i's believed worlds at w' if there is no i's preferred world at w where j tries to do α, that is, our aim is to avoid $G_i(w') \subseteq (B_i \circ R^{att}_{j:\alpha})(w)$ if $(R^{att}_{j:\alpha} \circ G_i)(w) = \emptyset$. Thus, for any $i, j \in AGT$, $\alpha \in AGT$ and $w \in W$ we have that:

4C. if $(R^{att}_{j:\alpha} \circ G_i)(w) \neq \emptyset$ then
$(G_i \circ R^{att}_{j:\alpha})(w) \subseteq (B_i \circ R^{att}_{j:\alpha})(w) \cup (R^{att}_{j:\alpha} \circ G_i)(w)$.

4.4.1. *Discussion*

In Herzig & Longin's logic of action and intention (2004) the NL and NF axioms are used for modelling goal dynamics. The following two axioms are accepted as valid principles describing how goals of agents change over time. For any $i, j \in AGT$ and $\alpha \in AGT$:

- $Goal_i\,[j:\alpha]\,\psi \rightarrow [j:\alpha]\,Goal_i\psi \vee Goal_i\,[j:\alpha]\,\bot$
- $[j:\alpha]\,Goal_i\psi \rightarrow Goal_i\,[j:\alpha]\,\psi \vee [j:\alpha]\,\bot$

where $[i:\alpha]$ is the standard operator of dynamic logic which is similar to our $[\![i:\alpha]\!]$. It has to be noted that these two axioms are based on an assumption of temporal consistency of choices. Our interest here is in clarifying the meaning of this assumption by investigating what we would deduce in $\mathcal{LIA}$ if the NL and NF axioms were introduced to model goal dynamics. To this aim, we focus on the following scenario.

> "Ulysses is going back to the island of Ithaca after the Trojan war. He knows that the morning after he will pass with his ship by the Island of the Sirens. Ulysses knows that no human has ever resisted the bewitching voice of the sirens: if someone passes by the island with his ship and hears the voice of the sirens, he is so attracted by this sound that he cannot prevent himself from desiring to steer his ship fatally into the rocks where the sirens are singing. Ulysses wants to refrain from steering his ship into the rocks since he does not want to die. But he is so curious to hear the voice of the sirens that he decides to pursue the following strategy. He decides to try to tie himself to the mast before approaching the island of the sirens so that, when he hears the voice of the sirens, he will not steer his ship into the rocks.

It can be proved that, under the NL and NF axioms for chosen goals, the scenario of Ulysses contains an inconsistency. Indeed, if either the NF Axiom for goals of the form $Goal_i [\![j : \alpha]\!] \varphi \rightarrow [\![j : \alpha]\!] Goal_i \varphi \vee Goal_i [\![j : \alpha]\!] \bot$ or the NL Axiom for goals of the form $[\![j : \alpha]\!] Goal_i \psi \rightarrow Goal_i [\![j : \alpha]\!] \psi \vee [\![j : \alpha]\!] \bot$ is supposed, it can proved that the two formulas $Goal_{Ulysses} \langle\!\langle Ulysses : tieMast \rangle\!\rangle \neg steerShip$ and $Bel_{Ulysses} \langle\!\langle Ulysses : tieMast \rangle\!\rangle Goal_{Ulysses} steerShip$ are inconsistent. This means that, under the NF (or NL) Axiom for chosen goals, Ulysses cannot want to try to tie himself to the mast in order to refrain from steering his ship into the rocks afterwards, when he expects that he will want to steer his ship into the rocks after he tries to tie himself to the mast. As the example of Ulysses and the sirens shows, if we supposed the NL or the NF axiom for chosen goals we would place a very strong constraint on goal dynamics. In fact, these two axioms require that chosen goals (i.e. choices) of an agent are always temporally consistent and an agent cannot expect that after the occurrence of a certain event he will want φ to be true when he actually wants φ to be false after the occurrence of such an event.[7] Here, we weaken the NF and NL axioms for chosen goals in order to be able to model agents who do not necessarily have temporally stable choices and goals. Thus, we improve over Herzig and Longin (2004). Our weaker version of a NF axiom for chosen goals is expressed by Axiom 4c. Under this refined version of a NF axiom for chosen goals the scenario of Ulysses is not problematic anymore. Indeed, under Axiom 4c, an agent may expect that after trying to do action α he will want φ to be true when at present he wants φ to be false after he tries to do α. More generally, in our logic choices of agents are not always stable. For instance, in our logic the following formula is consistent: $Goal_i \langle\!\langle j : \alpha \rangle\!\rangle \varphi \wedge [\![j : \alpha]\!] Goal_i \neg \varphi$.

4.5. Intentional attempt

The complementary Axioms 5c and 6c are rather new in the logical literature on action and intention. The formulas corresponding to these axioms are closed formulas, that is, formulas which contain no propositional letters (Blackburn et al., 2001). The semantic constraints which correspond to Axiom 5c and Axiom 6c are the following. For any $i \in AGT$, $\alpha \in AGT$ and $w \in W$:

5C. if $R^{att}_{i:\alpha}(w) \neq \emptyset$ then $\forall w' \in G_i(w)$, $R^{att}_{i:\alpha}(w') \neq \emptyset$.

6C. if $R^{att}_{i:\alpha}(w) = \emptyset$ then $\forall w' \in G_i(w)$, $R^{att}_{i:\alpha}(w') = \emptyset$.

According to Axiom 5c, an agent tries to do α only if he has a chosen goal to try to do α here and now. We suppose that $Goal_i\langle\langle i : \alpha\rangle\rangle\top$ corresponds to i's proximal intention to do α (see section 6.1 for a discussion). Thus, Axiom 5c can also be read, " agent i tries to do α only if he has the proximal intention to do α". This Axiom is an approximation of Principle 2 discussed in section 2.3 and is justified by the fact that in our logic any α which appears in a formula $\langle\langle i : \alpha\rangle\rangle\top$ is a basic action type of agent i (i.e. a bodily movement in i's repertoire). According to Principle 2, given a basic action type α of agent i, i's attempt to perform α is necessarily caused by i's proximal intention to perform α. As noted in section 2.3, it is improper to say that an agent i tries/attempts to perform a certain movement α (i.e. "i exerts his voluntary control over the execution of α"/"i goes through the mental effort of performing movement α") when this alleged attempt corresponds to a non-intentional form of behavior such as a spontaneous reaction (e.g. a reflex) or a reactive response to a given stimulus.

Axiom 6c which can be rewritten as $\neg Goal_i [\![i : \alpha]\!]\bot \rightarrow \langle\langle i : \alpha\rangle\rangle\top$ establishes that if an agent does not want to refrain from trying α now (i.e. he has no reason for not trying to do α now), then he tries to do α. This axiom is justified by: 1) the instance $\neg Goal_i [\![i : \alpha]\!]\bot \rightarrow Bel_i\neg Goal_i [\![i : \alpha]\!]\bot$ of axiom 6b (negative introspection of chosen goals) according to which: if i has no reason for not trying to do α now then he believes this; 2) an assumption about motivational power of beliefs according to which: i's belief that he has no reason for not trying to do α now (i.e. $Bel_i\neg Goal_i [\![i : \alpha]\!]\bot$) constitutes i's good reason for trying to do α now which pushes i to try to do α (i.e. $\langle\langle i : \alpha\rangle\rangle\top$).[8]

4.6. Soundness and completeness

We call $\mathcal{LIA}$ the logic axiomatized by the Axioms 0, 1a-8a, 1b-7b, 1c-6c and rules of inference R1-R5 given above. We call $\mathcal{LIA}$ *models* the set of those models in $\mathbf{M}$ satisfying all the semantic constraints 2A-4A,

6A-8A, 2B-3B, 5B-7B, 1C-6C that we introduced in sections 4.1-4.5. (Axioms 0, 1a, 5a, 1b, 4b, and R1-R5 don't require constraints because our semantics is that of a normal modal logic: the constraints are 'built' into Kripke models). We write $\vdash_{\mathcal{LIA}} \varphi$ if formula φ is a theorem of $\mathcal{LIA}$, i.e. if there is a proof of φ from Axioms 0, 1a-8a, 1b-7b, 1c-6c and rules of inference R1-R5. Moreover, $\models_{\mathcal{LIA}} \varphi$ denotes the fact that formula φ is *valid* in all $\mathcal{LIA}$ models, i.e. $M, w \vDash \varphi$ for every $\mathcal{LIA}$ model M and world w in M. Finally, a formula φ is said to be *satisfiable* if there is a $\mathcal{LIA}$ model M and some world w in M such that $M, w \vDash \varphi$. Now, we can prove that $\mathcal{LIA}$ is *sound* and *complete* with respect to the class of models satisfying all the semantic constraints imposed in sections 4.1-4.5. Formally:

THEOREM 1. *$\vdash_{\mathcal{LIA}} \varphi$ if and only if $\models_{\mathcal{LIA}} \varphi$.*

Proof of Theorem 1. All axioms of time and attempt (1a-8a), all axioms of mental attitudes (1b-7b) and all axioms of mental attitude dynamics (1c-6c) are in the Sahlqvist class, for which a general algorithm to compute their semantic counterparts exists. Therefore it is a routine task to verify that each axiom in 1a-8a, 1b-7b and 1c-6c corresponds to the respective semantic property described in sections 4.1-4.5. A general completeness result exists for all axioms which are in the Sahlqvist class (see Blackburn et al., 2001).

5. Properties of attempts

The aim of this section is to provide a formal relationship between attempts and mental attitudes on the one hand (section 5.1), and between attempts and action occurrences on the other hand (section 5.2).

5.1. Attempt and mental attitudes

The following theorems highlight the bridging role of attempt between mental level and executive level.

THEOREM 2. *For any $i \in AGT$ and $\alpha \in ACT$*

1. $\vdash_{\mathcal{LIA}} \langle\!\langle i : \alpha \rangle\!\rangle \top \leftrightarrow Goal_i \langle\!\langle i : \alpha \rangle\!\rangle \top$
2. $\vdash_{\mathcal{LIA}} [\![i : \alpha]\!] \bot \leftrightarrow Goal_i [\![i : \alpha]\!] \bot$
3. $\vdash_{\mathcal{LIA}} \langle\!\langle i : \alpha \rangle\!\rangle \top \leftrightarrow Bel_i \langle\!\langle i : \alpha \rangle\!\rangle \top$
4. $\vdash_{\mathcal{LIA}} [\![i : \alpha]\!] \bot \leftrightarrow Bel_i [\![i : \alpha]\!] \bot$

5. $\vdash_{\mathcal{LIA}} Goal_i\langle\!\langle i:\alpha\rangle\!\rangle\top \leftrightarrow Bel_i\langle\!\langle i:\alpha\rangle\!\rangle\top$

6. $\vdash_{\mathcal{LIA}} Goal_i\,[\![i:\alpha]\!]\,\bot \leftrightarrow Bel_i\,[\![i:\alpha]\!]\,\bot$

We only discuss Theorem 2.1, 2.3 and 2.4. Theorem 2.1 accounts for the conditions for passing from a mental state called intention to the executive and physical reality passing through a mental process called attempt. It says that an agent tries to perform movement α if and only if he has the proximal intention to perform this movement. The direction $\langle\!\langle i:\alpha\rangle\!\rangle\top \rightarrow Goal_i\langle\!\langle i:\alpha\rangle\!\rangle\top$ of Theorem 2.1 is Axiom 5c. As noted in section 4.5, this axiom should be conceived as a formal translation of Principle 2 given in section 2.3 which says that: if α is a basic action type of agent i then, i's attempt to perform α is necessarily caused by i's proximal intention to peform α. The other way round, the direction $Goal_i\langle\!\langle i:\alpha\rangle\!\rangle\top \rightarrow \langle\!\langle i:\alpha\rangle\!\rangle\top$ of Theorem 2.1 should be conceived as a formal translation of Principle 1 also given in section 2.3 which says that: whenever an agent has a present-directed intention to perform α and α is a basic action type of the agent, such an intention brings about the agent's attempt to do α. However, it has to be noted that Theorem 2.1 is just a approximation of the causal relation between proximal intention and attempt as specified by Principles 1 and 2. Indeed, our logic $\mathcal{LIA}$ is not sufficiently expressive to capture a true notion of causality and can only "simulate" the causal relation between proximal intention and attempt by stating that the former implies the latter and the latter implies the former.

According to Theorem 2.3, if an agent tries to perform movement α then he believes this and vice versa, that is, an agent is always aware of the bodily movement is engaged in performing. For example, while raising an arm, an agent is aware of the fact that is in the process of moving his arm. This means, while performing a certain bodily movement, an agent is aware of the actual state of his motor systems. This property of volitional attempt is accepted by other authors (Davis, 1979, Ginet, 1990).[9]

5.2. Attempts and basic action tokens

As far as *attempt* and *action* are concerned, the following general Principle 3 is a fundamental basis for understanding the relationship between these two concepts.

PRINCIPLE 3. *Given a basic action type α of agent i: 1) if the execution precondition of α for agent i holds and i tries to do α, then i's attempt will be successful; 2) if the execution precondition of α for agent i does not hold and i tries to do α, then i's attempt will fail.*

According to Principle 3, given a basic action type α of agent i, α is successfully performed by i if and only if i tries to do α and the execution precondition of α for agent i holds. This principle can be formally specified in $\mathcal{LIA}$. To this end, new formal constructs are introduced. We call *closed attempt formulas* constructions of the form $\langle\langle i : \alpha\rangle\rangle\top$ and we denote with Δ the set of *closed attempt formulas*, that is, $\Delta = \{\langle\langle i : \alpha\rangle\rangle\top | i \in AGT, \alpha \in ACT\}$. From Δ and the set of atomic formulas Π, the set of *objective formulas* OBJ is defined as follows.

DEFINITION 1. *OBJ is the smallest superset of Π and Δ such that: if $\varphi, \psi \in OBJ$ then $\neg\varphi, \varphi \vee \psi \in OBJ$*

We suppose that Pre is a function which assigns an objective formula in OBJ to each basic action type of each agent, that is: $Pre : ACT \times AGT \longrightarrow OBJ$, where formula $Pre(i, \alpha)$ denotes the *execution precondition* of α for agent i.[10] The following abbreviation is given for any $i \in AGT$ and $\alpha \in AGT$ in order to express the concept of successful execution of a basic action type (or basic action token).

DEFINITION 2. *Successful execution of a basic action type (or basic action token).*
$\langle i : \alpha\rangle^s\varphi =_{def} \langle\langle i : \alpha\rangle\rangle\varphi \wedge Pre(i, \alpha)$

According to Definition 2, i performs α in a successful way and φ holds after α's occurrence if and only if i tries to perform α, φ holds after this attempt of i and, the execution precondition of α for i holds. As the following instance of Definition 2 shows, we are finally able to provide a formal translation of Principle 3 about the relation between *attempt* and *successful performance of a basic action*: $\vdash_{\mathcal{LIA}} \langle i : \alpha\rangle^s\top \leftrightarrow \langle\langle i : \alpha\rangle\rangle\top \wedge Pre(i, \alpha)$. According to this, a given basic action type α of agent i is *successfully* performed by i if and only if i tries to do α and the execution precondition of action α for i holds. This equivalence should be conceived as a non-standard way to express *execution laws*. In fact, *execution laws* have been traditionally expressed by taking actions as primitive elements, without decomposing them into more elementary constituents (viz. attempts) (Castilho et al., 1999, Reiter, 2001). For instance, in Reiter (2001) it is supposed that an action α is executable if and only if its execution precondition $Poss(\alpha)$ is true. Thus, in Reiter's approach the concept of attempt only appears implicitly in the concept of execution precondition and there is no clear distinction between the former and the latter. On the contrary, in our approach *attempt* and *execution precondition* are clearly distinguished in the formal specification of execution laws.

The distinction between attempt and execution precondition is also crucial for defining the concept and corresponding operator $\langle i : \alpha\rangle^f$

of *action failure* in a simple way: $\langle i:\alpha\rangle^f\varphi =_{def} \langle\langle i:\alpha\rangle\rangle\varphi \wedge \neg Pre(i,\alpha)$. Thus, agent i fails to perform a basic action α and φ holds after i's failure (i.e. $\langle i:\alpha\rangle^f\varphi$) if and only if i tries to perform α φ holds after i's attempt to perfom α and, the execution precondition of α for i does not hold. Note that the dual of $\langle i:\alpha\rangle^s$ and the dual of $\langle i:\alpha\rangle^f$ can be defined according to the following abbreviations: $[i:\alpha]^s\varphi =_{def} \neg\langle i:\alpha\rangle^s\neg\varphi$, $[i:\alpha]^f\varphi =_{def} \neg\langle i:\alpha\rangle^f\neg\varphi$, where $[i:\alpha]^s\varphi$ is read "after i performs α in a successful way, it is the case that φ" and $[i:\alpha]^f\varphi$ is read "after i fails to perform α, it is the case that φ". From the previous definitions of action success and action failure the following two theorems are derivable: $\vdash_{\mathcal{LIA}} \langle\langle i:\alpha\rangle\rangle\varphi \leftrightarrow \langle i:\alpha\rangle^s\varphi \vee \langle i:\alpha\rangle^f\varphi$ and $\vdash_{\mathcal{LIA}} [\![i:\alpha]\!]\varphi \leftrightarrow [i:\alpha]^s\varphi \wedge [i:\alpha]^f\varphi$. This shows that the concept of attempt on the one hand and the concepts of action success and action failure on the other hand are interdefinable.

5.2.1. *Discussion and example*

Now that we have given formal characterizations of the notion of execution precondition, action success and action failure, a few comments are in order. First, we want to emphasize that our choice to have an agent argument for the function Pre has a precise rationale. In fact, as clarified in section 3.1, attempts to do the same basic action α of different agents i and j are supposed to be distinct events in our logic (e.g. "Bob's attempt to move his leg" is different from "Jack's attempt to move his leg"). Thus, the conditions under which i's attempt to perform α is successful might be different from the conditions under which j's attempt to perform α is successful, that is, the execution precondition of α for i might be different from the execution precondition of α for j (e.g. the execution precondition of "moving a leg" is $BobFreeLeg \wedge \neg BobParalyzedLeg$ for Bob and $JackFreeLeg \wedge \neg JackParalyzedLeg$ for Jack).

As regards the ontological status of *execution precondition*, it has to be noted that according to Definition 2 $Pre(i,\alpha)$ denotes both: (a) the necessary conditions for agent i's successful performance of action α; (b) the conditions which are, together with the fact that agent i tries to do α, sufficient conditions for agent i's successful performance of action α. In particular, formula $Pre(i,\alpha)$ expresses the following two facts. (1) There are no external forces such as actions and attempts of other agents and physical obstacles which can interfere with an agent i's performance of a basic action α and can make it physically impossible for i to perform α in a successful way. (2) There are no impairments,

Figure 1. Example.

mental deficiencies, disabilities, etc. which prevent i from moving his body.

By way of example, suppose that there are two robots $r1$ and $r2$ living in the house with two rooms represented in Fig. 1. The two robots have $\Rightarrow$ (move right) and $\Leftarrow$ (move left) in their repertoires of bodily movements. The set of atomic formulas $\{r1L, r1R, r2L, r2R\}$ encode the positions of a robot in the environment. Namely, $r1L$ (viz. $r2L$) means that robot 1 (viz. robot 2) is in the left room, $r1R$ (viz. $r2L$) means that robot 1 (viz. robot 2) is in the right room. The rules of the game say that if one of the two robots is in the left room then he can only move right and, if one of the two robots is in the right room then he can only move left. Indeed, a wall surrounds the house and a robot cannot climb over the wall by moving right (viz. left) when he is in the right (viz. left) room. Moreover, if the two robots try to move at the same time in order to pass through the door, both attempts will fail and the robots' positions will not change. Indeed, the door between the two rooms is so narrow that robot 1 and robot 2 cannot pass together through it. For example, suppose that $r1L$ and $r2R$ hold and $r1$ tries to do $\Rightarrow$. Then $r1$ will succeed in performing $\Rightarrow$ if and only if $r2$ does not try to do $\Leftarrow$. Now, suppose that $r1L$ and $r2L$ hold and $r1$ tries to do $\Rightarrow$. Then $r1$ will succeed in performing $\Rightarrow$ if and only if $r2$ does not try to do $\Rightarrow$. The rules of the game can be specified by our function Pre as follows:

$Pre(r1, \Leftarrow) = r1R \wedge (r2R \rightarrow [\![r2 :\Leftarrow]\!] \bot) \wedge (r2L \rightarrow [\![r2 :\Rightarrow]\!] \bot)$;
$Pre(r2, \Leftarrow) = r2R \wedge (r1R \rightarrow [\![r1 :\Leftarrow]\!] \bot) \wedge (r1L \rightarrow [\![r1 :\Rightarrow]\!] \bot)$;
$Pre(r1, \Rightarrow) = r1L \wedge (r2L \rightarrow [\![r2 :\Rightarrow]\!] \bot) \wedge (r2R \rightarrow [\![r2 :\Leftarrow]\!] \bot)$;
$Pre(r2, \Rightarrow) = r2L \wedge (r1L \rightarrow [\![r1 :\Rightarrow]\!] \bot) \wedge (r1R \rightarrow [\![r1 :\Leftarrow]\!] \bot)$.

6. Dynamics of intentions

After having investigated the concept of attempt and its relation with the concept of action, we will look in this section at dynamics of intentions. We will define achievement goals, future-directed intentions and present-directed intentions. Then, we will show how in our logic it is possible to analyze *generation* of future-directed intentions through in-

strumental reasoning. Finally, we will investigate the issue of *persistence* of intentions.

As a preliminary step toward a formal analysis of intention dynamics, we provide a formal definition of *achievement goal* for any $i \in AGT$.

DEFINITION 3. *Achievement goal.*
$AGoal_i\varphi =_{def} Goal_i\Diamond Bel_i\varphi \wedge \neg Bel_i\varphi$

The previous definition is different from the Cohen and Levesque (1990) definition of achievement goal: $AGoal_i^{CL}\varphi =_{def} Goal_i\Diamond\varphi \wedge Bel_i\neg\varphi$ and it is identical to the one proposed in Herzig and Longin (2004). Agent i has an achievement goal that φ if and only if he is motivated to achieve a future state at which he believes that φ holds. Thus, the kind of achievement goal that we define in this paper is rather an "epistemized" kind of achievement goal: $\vdash_{\mathcal{LIA}} AGoal_i\varphi \leftrightarrow AGoal_i^{CL}Bel_i\varphi$.

6.1. Defining distal and proximal intentions

We suppose that a *future-directed intention to do* (or *distal intention to do*) is a future-oriented chosen goal whose content is a future basic action of the agent. The content of a *future-directed intention to do* α can be either the future successful execution of α or simply a future attempt to do α. For instance, we assume that the expression "agent i intends to raise an arm later" is the vehicle of one of the following two meanings: either "i intends to raise an arm later in a successful way" or "i intends to try to raise an arm later". Indeed, when i intends to raise an arm later, he may simply have a mental representation of himself undertaking the bodily movement of raising an arm independently from the fact that the arm is going to be raised. In this case, the agent has a mental prospect of the executive part of the basic action of raising an arm. The mental representation "i undertaking the action of raising an arm" is in the content of i's intention to try to raise an arm later. But i's intention to raise an arm later could be more complex. Indeed, when he intends to raise an arm later, i may have the mental representation of himself undertaking the bodily movement of raising an arm and succeeding in raising the arm. In this case, the agent has a mental prospect of both the executive part of the basic action of raising an arm and the intrinsic result of that action (i.e. the result "the arm is raised"). The mental representation "i undertaking the action of raising an arm and i succeeding raising the arm" is in the content of i's intention to raise an arm later in a successful way. Generally speaking, we suppose that a *future-directed intention to do* α can be either a

future-directed intention to do α in a successful way or a *future-directed intention to try to do α*.

The previous concepts can be formalized in $\mathcal{LIA}$. With "agent i has a future-directed intention to do α in a successful way" we mean "agent i has the chosen goal to do α later and he is not trying to do α yet": $FDI_i\langle i:\alpha\rangle^S\top =_{def} Goal_i\Diamond\langle i:\alpha\rangle^S\top \wedge \neg\langle\langle i:\alpha\rangle\rangle\top$,
where the second element of the conjunction is given in order to exclude the present-directed intention to do α. With "agent i has a future-directed intention to try to do α" we mean "agent i has the chosen goal to try to do α later and agent i is not trying to do α yet": $FDI_i\langle\langle i:\alpha\rangle\rangle\top =_{def} Goal_i\Diamond\langle\langle i:\alpha\rangle\rangle\top \wedge \neg\langle\langle i:\alpha\rangle\rangle\top$. Finally, with "agent i has a future-directed intention to do α" we mean "agent i has a future-directed intention to do α in a successful way or agent i has a future-directed intention to try to do α": $FDI_i(\alpha) =_{def} FDI_i\langle\langle i:\alpha\rangle\rangle\top \vee FDI_i\langle i:\alpha\rangle^S\top$. Given Definition 2 of *successful execution of a basic action type*, we can easily deduce the following property: $\vdash_{\mathcal{LIA}} FDI_i\langle i:\alpha\rangle^S\top \rightarrow FDI_i\langle\langle i:\alpha\rangle\rangle\top$. Thus, we can conclude that $FDI_i(\alpha)$ is equivalent to $FDI_i\langle\langle i:\alpha\rangle\rangle\top$. This means that in $\mathcal{LIA}$ an *intention to do* reduces to an *intention to try to do*. Thus, for any $i \in AGT$ and $\alpha \in ACT$ the following abbreviation is given.

DEFINITION 4. *Future-directed intention (or distal intention).*
$FDI_i(\alpha) =_{def} Goal_i\Diamond\langle\langle i:\alpha\rangle\rangle\top \wedge \neg\langle\langle i:\alpha\rangle\rangle\top$

Now, we can prove that a future-directed intention to do α is equivalent to an achievement goal of trying to do α, that is:
$\vdash_{\mathcal{LIA}} FDI_i(\alpha) \leftrightarrow AGoal_i\langle\langle i:\alpha\rangle\rangle\top$. Besides, as the following theorem shows, we can prove that agents are "correctly aware" of their future-directed intentions: $\vdash_{\mathcal{LIA}} FDI_i(\alpha) \leftrightarrow Bel_iFDI_i(\alpha)$. We conclude by pointing out that the formula $FDI_i\langle\langle i:\alpha\rangle\rangle\top \rightarrow FDI_i\langle i:\alpha\rangle\top$ is not valid. Thus, in our logic an agent may intend to try to do α without intending to do α in a successful way, i.e. we can find a $\mathcal{LIA}$ model in which $FDI_i\langle\langle i:\alpha\rangle\rangle\top \wedge \neg FDI_i\langle i:\alpha\rangle\top$ is satisfiable. The plausibility of such a consequence has been defended by Bratman, Bratman, Mele (1984, 1987, 1992) and criticized by McCann (1986). In our view this consequence is acceptable. In fact, we can imagine plenty of scenarios where an agent intends to try to do something without intending to accomplish it. For example, suppose that i promises to pay j a certain amount of euros if j tries to raise an arm within five seconds. i assures j that he need not actually raise an arm in a successful way for getting the amount of euros. It is plausible to say that j intends to try to raise an arm even if he does not intend to succeed in raising an arm. In fact, j does not care whether his trying is going to succeed or to fail.

Similarly to future-directed intentions, in $\mathcal{LIA}$ we can define two kinds of *present-directed intention to do* (or *proximal intention to do*) depending on the content of the motivational attitude. We suppose that the expression "agent i intends to do α now" means either "agent i intends to do α now in a successful way" or "agent i intends to try to do α now". The former mental attitude is captured by the formula $Goal_i\langle i : \alpha\rangle^s\top$ whereas the latter mental attitude is captured by the formula $Goal_i\langle\langle i : \alpha\rangle\rangle\top$. The previous argument for future-directed intentions applies to present-directed intentions as well. Indeed, since "agent i has the present-directed intention to do α in a successful way" (i.e. $Goal_i\langle i : \alpha\rangle^s\top$) always implies that "agent i has the present-directed intention to try to do α" (i.e. $Goal_i\langle\langle i : \alpha\rangle\rangle\top$), and "agent i has the present-directed intention to do α" means "agent i has the present-directed intention to do α in a successful way or agent i has the present-directed intention to try to do α" (i.e. $Goal_i\langle\langle i : \alpha\rangle\rangle\top \vee Goal_i\langle i : \alpha\rangle^s\top$), we conclude that a present-directed intention to do α reduces to a present-directed intention to try to do α. This argument leads to the following definition for any $i \in AGT$ and $\alpha \in ACT$.

DEFINITION 5. *Present-directed intention (or proximal intention).* $PDI_i(\alpha) =_{def} Goal_i\langle\langle i : \alpha\rangle\rangle\top$

6.2. Instrumental reasoning

Practical reasoning is commonly conceived as reasoning that concludes in an action or in an intention. To avoid confusion some philosophers have distinguished between *practical reasoning* (or *instrumental reasoning*) and *practical argument* (Audi, 1982). The former designates a process of passing from appropriate premises (a superior chosen goal or intention and an instrumentality belief) to a practical conclusion that is aimed at action (the instrumental intention), whilst the latter designates the corresponding structure of propositions (premises + practical conclusion). In this section of the paper we will show that a practical argument is formally derivable in our logic. As far as we know, there is no logical theory of action and intention which has tried to rebuild practical reasoning and the corresponding practical arguments inside a logic of mental states. Some authors resolve the problem by hand (Dignum and Conte, 1998, Meyer et al., 1999, Panzarasa et al., 2002) by *assuming* that certain practical arguments are valid principles of intention generation.[11] The main weakness of such approaches is that intention generation and practical reasoning are not really properties of the logic in question. More precisely, in such approaches practical arguments are neither proved to be logical consequences of the axioms

and inference rules of the logic, nor are they taken as axioms or inference rules and analyzed at the semantic level.

Several forms of practical reasoning and corresponding practical arguments have been studied in philosophy. An interesting variant is a form of *necessity-based practical reasoning* where conclusions drawn always concern *necessary* means to some end. Von Wright (1972) characterizes *necessity-based practical reasoning* by a practical argument of the following form: *i has the chosen goal that φ will be achieved* (or *i intends to make it true that φ*) and *believes that he will not achieve φ unless he does α*, therefore *i should intend to do α*. Some refinements of this form of necessity-based practical reasoning and the related practical argument have been proposed (Bratman, 2005, Broome, 2002, Wallace, 2001). For instance, Bratman, Wallace (2005, 2001) have claimed that it is only when an agent believes his present decision to do a certain action is needed to achieve his superior goals that he should form the appropriate intention to do the action in question. Thus, in their perspective, necessity-based practical reasoning must be characterized by a practical argument with the following structure: *i has the chosen goal that φ will be achieved* (or *i intends to make it true that φ*) and *believes that he will achieve φ only if he now forms the intention to do α*, therefore *i should intend to do α*. In $\mathcal{LIA}$ this kind of practical argument can be explicitly captured and somehow refined. This is what the following theorems highlight.

THEOREM 3. *For any $i \in AGT$ and $\alpha \in ACT$*

1. $\vdash_{\mathcal{LIA}} Bel_i(\Diamond Bel_i\varphi \rightarrow FDI_i(\alpha)) \wedge AGoal_i\varphi \rightarrow FDI_i(\alpha)$
2. $\vdash_{\mathcal{LIA}} Bel_i(\Diamond Bel_i\varphi \rightarrow PDI_i(\alpha)) \wedge AGoal_i\varphi \rightarrow PDI_i(\alpha)$

According to theorem 3.1 if agent i has an achievement goal that φ and he believes that he will reach his achievement goal that φ only if he now forms the intention to do α in the future then, he should form the future-directed intention to do α. This is typical of situations in which, according to the agent's beliefs, the decision to do α cannot be delayed. In these situations the agent behaves as if he wrote down something in his mental agenda in order to remember that he has to do α. Imagine Bill planning to give a party withing few days. He has the achievement goal that his friend Jack will be at the party: $AGoal_{Bill}JackAtParty$. Furthermore, he believes that he will reach his achievement goal only if he now forms the intention to ask Jack to come to the party: $Bel_{Bill}(\Diamond Bel_{Bill}JackAtParty \rightarrow FDI_{Bill}(ask))$. Indeed, according to Bill, the decision to ask Jack cannot be delayed since taking the decision in advance is the only way to ensure that he will (remember to) ask Jack to come to the party when he will meet

him. Therefore (according to Theorem 3.1) Bill should form the future-directed intention to ask Jack to come to the party: $FDI_{Bill}(ask)$. Theorem 3.2 concerns generation of present-directed intentions through practical reasoning. According to this theorem, if agent i has an achievement goal that φ and he believes that he will reach his achievement goal that φ only if he now forms the intention to immediately do α then, he should form the proximal intention to do α.

6.3. Intention and expectation of a future attempt

Before dealing with the problem of intention *persistence*, we will briefly analyze the expectation which is involved in every future-directed intention. The following theorem shows that a future-directed intention to do α implies an expectation of a future attempt to do α.

THEOREM 4. *For any* $i \in AGT$ *and* $\alpha \in ACT$
$\vdash_{\mathcal{LIA}} FDI_i(\alpha) \rightarrow Bel_i \Diamond \langle\!\langle i : \alpha \rangle\!\rangle \top$

According to Theorem 4 if an agent has a future-directed intention to do α then he believes that he will try to do α. Due to the equivalence between $PDI_i(\alpha)$ and $\langle\!\langle i : \alpha \rangle\!\rangle \top$ (Theorem 2.1 and Definition 5), the following formula is a theorem of $\mathcal{LIA}$ as well: $\vdash_{\mathcal{LIA}} FDI_i(\alpha) \rightarrow Bel_i \Diamond (PDI_i(\alpha) \wedge \langle\!\langle i : \alpha \rangle\!\rangle \top)$. This means that if an agent has the future-directed intention to do α then he believes that he will have the proximal intention to do α and (consequently) he will try to do α. This formal consequence is acceptable. Indeed, it seems quite reasonable to suppose that: if an agent intends to do α in the future then he believes that, if things go as expected, due to this intention he will try to do α at will, that is, he believes that if things go as expected then his future-directed intention to do α will persist until it transforms into a present-directed intention and leads him to try to do α. Thus, an intention to do α involves a sort of self-referential aspect: the belief that such an intention to do α will be responsible for a future attempt to do α. Also Harman (1986) and Searle (1983) have emphasized this self-referential nature of intention.

The previous two theorems of $\mathcal{LIA}$ do not say anything about the expectation of the future *successful* execution of an intended action. Indeed, the following implication is not valid in $\mathcal{LIA}$: $FDI_i(\alpha) \rightarrow Bel_i \Diamond \langle i : \alpha \rangle \top$. Thus, in $\mathcal{LIA}$ it is not always the case that if an agent intends to do α in the future, then he believes that he will succeed in performing α. We agree with Bratman, McCann (1987, 1991) that such a relationship between intentions and beliefs would be too strong. Note that the following implication is not valid either: $Bel_i \Diamond \langle\!\langle i : \alpha \rangle\!\rangle \top \rightarrow FDI_i(\alpha)$. Thus, in our framework a belief of a future attempt to do

α does not necessarily imply a future-directed intention to do α. We have reason to claim that this implication would be too strong. Indeed, being sure that tomorrow morning I will (try to) wake up at 7 a.m. does not imply having already a future-directed intention to wake up at 7 a.m. tomorrow morning. The general claim is that there is no rational pressure to intend to do things that we expect to do in the future. Indeed, an intention has a motivational aspect that a mere belief does not have.

6.4. Persistence of Intentions

There are several authors who have taken an interest in the topic of *persistence of intentions* (Bratman, 1987, Cohen and Levesque, 1990, Van der Hoek et al., 2007). According to Bratman intentions have a certain kind of inertial force, that is, intentions tend to resist reconsideration. His idea is that the inertia of intentions is a crucial property of resource-bounded agents. The rationale behind this is that once an agent has settled himself to perform a certain action α, he cannot all the time spend energy to evaluate whether the chosen course of action is the right one. Once an agent has deliberated in favor of action α and has formed the intention to do α, he does not normally continue to deliberate whether doing α or not. The agent commits to performing action α and in the absence of relevant new information, the future-directed intention to do α will resist further reconsideration. According to Bratman, this qualifies intention for a functional role in cognition to which mere desires and wishes are unsuited. As the following theorem shows, in $\mathcal{LIA}$ such an inertial property of intentions can be captured.

THEOREM 5. *For any* $i, j \in AGT$ *and* $\alpha, \beta \in ACT$
$\vdash_{\mathcal{LIA}} FDI_i(\alpha) \wedge \neg Bel_i [\![j:\beta]\!] \bot \wedge \neg Goal_i [\![j:\beta]\!] \bot \rightarrow [\![j:\beta]\!] (FDI_i(\alpha) \vee \langle\!\langle i:\alpha \rangle\!\rangle \top)$

The reading of Theorem 5 is the following: if i has the future-directed intention to do α and it is compatible with his beliefs and goals that j is going to try to do β then after j tries to do β either 1) i has still the future-directed intention to do α or 2) i tries to do α. Due to the equivalence between $\neg Bel_i [\![i:\beta]\!] \bot \wedge \neg Goal_i [\![i:\beta]\!] \bot$ and $\langle\!\langle i:\beta \rangle\!\rangle \top$ (Theorems 2.2 and 2.4), in the case where $i = j$ Theorem 5 can be written in a more compact way: $\vdash_{\mathcal{LIA}} FDI_i(\alpha) \rightarrow [\![i:\beta]\!] (FDI_i(\alpha) \vee \langle\!\langle i:\alpha \rangle\!\rangle \top)$.[12] Therefore, if an agent has a future-directed intention to do α and is already engaged in doing action β, he does not reconsider his intention to do α. This means that in our logic an agent can abandon his future-directed intentions only if he is not busy doing something. Another instance of Theorem 5 is the following:

$\vdash_{\mathcal{LIA}} FDI_i(\alpha) \wedge \neg Bel_i [\![j:\alpha]\!] \bot \wedge \neg Goal_i [\![j:\alpha]\!] \bot \rightarrow [\![j:\alpha]\!] (FDI_i(\alpha) \vee \langle\!\langle i:\alpha \rangle\!\rangle \top)$.

This shows that in our logic the persistence of an agent i's future-directed intention to do α does not depend on the fact that a different agent j tries to do α, when j's attempt to do α is compatible with i's beliefs and goals. This consequence is acceptable given the main assumptions of this work. As emphasized in sections 3.1 and 5.2.1, i's attempt to do α and j's attempt to do α are different events in our logic which might produce different effects under the same circumstances. Therefore, it is reasonable to suppose that i's intention to move his body in a certain way α should not be affected by a different agent j moving his body in the same way α, when j's attempt to do α is compatible with i's beliefs (i.e. $\neg Bel_i [\![j:\alpha]\!] \bot$) and goals (i.e. $\neg Goal_i [\![j:\alpha]\!] \bot$). Since j's attempt to perform movement α is different from what i intends when he has the intention to perform movement α, the occurrence of j's attempt to perform α should be irrelevant for i's intention to perform α. The following two examples are given in order to justify such a property of an agent's intention. Imagine Fred working in a hot summer at his office and feeling thirsty. Hence, Fred forms the future-directed intention to walk to the soda machine in front of his office and take a bottle of water: $FDI_{Fred}(walk)$. It also compatible with Fred's goals and beliefs that his colleague Jack also tries to walk to the soda machine: $\neg Bel_{Fred} [\![Jack:walk]\!] \bot \wedge \neg Goal_{Fred} [\![Jack:walk]\!] \bot$. Now suppose that Jack tries to walk to the soda machine and Fred is informed of this fact. Fred's intention is not affected by Jack's attempt. Indeed, Fred has no reason to abandon his intention to walk to the soda machine after a different person has tried to do the same thing (Fred's thirst is not quenched by Jack's attempt to walk to the soda machine!). It seems reasonable to conclude, as the $\mathcal{LIA}$ Theorem prescribes, that after Jack tries to walk to the soda machine either Fred has still the future-directed intention to walk to the soda machine or he tries to do this: $[\![Jack:walk]\!] (FDI_{Fred}(walk) \vee \langle\!\langle Fred:walk \rangle\!\rangle \top)$. Now, suppose that Mary and her husband Bill have received a nice bunch of flowers from Bob as a present for the 20th anniversary of their marriage. At world w in a model M Mary wants she and Bill to be polite with Bob, that is, for every Mary's preferred history at w, there exists a future world in which $MaryPolite \wedge BillPolite$ holds: $\forall w' \in G_{Mary}(w), \exists v \in R_{\Box}(w')$ s.t. $M, v \models MaryPolite \wedge BillPolite$. Hence, Mary has the future-directed intention to tell Bob - Thank you Bob for the flowers!- : $FDI_{Mary}(thankBob)$. This means that, for every Mary's preferred history at w, there exists a future world in which Mary tries to thank Bob: $\forall w' \in G_{Mary}(w), \exists v \in R_{\Box}(w')$ s.t. $R^{att}_{Mary:thankBob}(v) \neq \emptyset$. Suppose that Bill's attempt to thank Bob fits in Mary's plans since she

thinks that Bill will be polite with Bob only if he personally thanks Bob. Therefore, it is compatible with Mary's goals that her husband Bill tries to thank Bob: $\neg Goal_{Mary}\,[\![Bill : thankBob]\!]\,\bot$. This means that $\exists w' \in G_{Mary}(w)$ s.t. $R^{att}_{Bill:thankBob}(w') \neq \emptyset$. Moreover, Mary considers it possible that Bill tries to thank Bob, that is, it is compatible with Mary's beliefs that that her husband Bill tries to thank Bob: $\neg Bel_{Mary}\,[\![Bill : thankBob]\!]\,\bot$. This means that $\exists w' \in B_{Mary}(w)$ s.t. $R^{att}_{Bill:thankBob}(w') \neq \emptyset$. Imagine at w Bill tries to thank Bob and Mary is informed of this. World v is the outcome of Bill's attempt at world w, i.e. $v \in R^{att}_{Bill:thankBob}(w)$. Mary's intention to thank Bob is not affected by Bill saying - Thank you Bob for the flowers!-. Indeed, there is no relation between Bob's attempt to thank Bill and the result $MaryPolite$ (i.e. $MaryPolite$ will still be false at v after Bill tries to thank Bob). This is reason why at v Mary should either persist with his future-directed intention to thank Bob or try to thank Bob (i.e. $FDI_{Mary}(thankBob) \vee \langle\!\langle Mary : thankBob \rangle\!\rangle \top$ should be true at v). Finally, let us slightly modify the scenario, and suppose that Mary considers it to be sufficient if either she or Bill thanks Bob. Moreover suppose that Mary prefers that she does it and Bill does not do it. Hence we have: $FDI_{Mary}(thankBob) \wedge Goal_{Mary}\,[\![Bill : thankBob]\!]\,\bot$ is true at w. In this case, Bill's attempt to thank Bob is unwanted by Mary and does not fit in her plans. She therefore revises her goals at world v after Bill tries to thank Bob (which is a process $\mathcal{LIA}$ does not account for), possibly dropping $FDI_{Mary}(thankBob)$.

7. Conclusion

A comprehensive formal model of intention, volitional attempt and action has been developed in this paper. We have focused on those kinds of actions called basic actions that correspond to bodily movements in the repertoire of an agent, that is, movements over which an agent can exert his voluntary control. We have investigated the relationships between attempts and basic action occurrences and between attempts (conceived as mental processes) and intentions (conceived as mental states). On the side of intention dynamics, we have debated two general issues: the issue of *intention generation* and *instrumental reasoning* and the issue of *intention persistence*. More generally, in this work the concept of volitional attempt has been integrated into a logical framework in which intention dynamics can be studied.

We hope that the analysis developed in this paper will be useful for improving understanding of intentional action and will offer an interesting perspective on logical modelling of intentional embodied

systems, that is, systems such as a (simulated) robot that intentionally interact with a simulated or real environment by the means of a defined, artificial body.

Notes

[1] See also Brand (1984) for a general overview of volitional theories of action.

[2] Note that non-basic actions should not be confused with sequences of basic actions (i.e. sequence of bodily movements under the voluntary control of an agent).

[3] The terms *future-directed intention* and *present-directed intention* are used by Bratman, whilst the terms *distal intention* and *proximal intention* are used by Mele. In this paper, the terms *future-directed intention* and *distal intention* on one side and *present-directed intention* and *proximal intention* on the other side are synonymous.

[4] A further distinction is between *intention-to do* and *intention-to be* (Bratman, 1987). The former involves the performance of some action (e.g. I have the intention to write a letter), whilst the latter involves the achievement of some state of affairs by performing some action (e.g. I intend to be in France tomorrow by catching the first train from Rome to Paris). In this paper we focus only on *intention-to do*. For a formalization of the concept of *intention-to be* see Cohen and Levesque, Grosz and Kraus (1990, 1996).

[5] Also Segerberg (1992) has emphasized that PDL can be exploited not only for reasoning about program executions but also for reasoning about action occurrences.

[6] More generally, we would have to introduce a function $Rep : AGT \longrightarrow 2^{ACT}$ where $Rep(i)$ denotes the set of basic action types of i (i.e. i's repertoire of bodily movements that can be under i's voluntary control) and to include in the language $\mathcal{L}_{\mathcal{LIA}}$ only formulas of type $\langle\langle i : \alpha \rangle\rangle \varphi$ where $\alpha \in Rep(i)$.

[7] See also Elster (1979) for a discussion on the issue of intertemporal choice applied to the "Ulysses and the sirens" problem.

[8] This should be conceived as a more general assumption about practical rationality according to which: "believing that I have no reason for refraining from doing α now constitutes a good reason for me to do α now".

[9] Nevertheless there are interesting works in neuroscience (Frith et al., 2000) showing that such a capacity is impaired in humans with damages to neural substrates of the motor system.

[10] Given that time is linear in $\mathcal{LIA}$, we prefer using the term *execution precondition* instead of the term *executability precondition*.

[11] Other authors have resolved the problem by introducing external components such as conditional planning rules (Broersen et al., 2002, Thomason, 2000).

[12] Note that under the conditions $i = j$ and $\alpha = \beta$ the theorem can be proved in a trivial way. In fact, $FDI_i(\alpha)$ implies $[\![i : \alpha]\!] \bot$ (by Definition 4). From this, we can easily conclude that $FDI_i(\alpha) \rightarrow [\![i : \alpha]\!] (FDI_i(\alpha) \vee \langle\langle i : \alpha \rangle\rangle \top)$ is a theorem of $\mathcal{LIA}$.

References

Audi, R.: 1982, 'A theory of practical reasoning'. *American Philosophical Quarterly* **19**, 25–39.

Baltag, A., L. Moss, and S. Solecki: 1998, 'The Logic of Public Announcements, Common Knowledge and Private Suspicions'. In: *Proc. Seventh Conference on Theoretical Aspects of Rationality and Knowledge (TARK'98)*. San Francisco, CA, pp. 43–56, Morgan Kaufmann.

Belnap, N., M. Perloff, and M. Xu: 2001, *Facing the future: agents and choices in our indeterminist world*. New York: Oxford University Press.

Blackburn, P., M. de Rijke, and Y. Venema: 2001, *Modal Logic*. Cambridge: Cambridge University Press.

Brand, M.: 1984, *Intending and Acting*. Cambridge: MIT Press.

Bratman, M.: 1984, 'Two Faces of Intention'. *Philosophical Review* **93**, 375–405.

Bratman, M.: 1987, *Intentions, plans, and practical reason*. Cambridge: Harvard University Press.

Bratman, M.: 2005, 'Intention, belief, practical, theoretical'. *Unpublished manuscript*.

Broersen, J., M. Dastani, J. Hulstijn, and L. van der Torre: 2002, 'Goal Generation in the BOID architecture'. *Cognitive Science Quarterly* **2(3-4)**, 428–447.

Broome, J.: 2002, 'Practical reasoning'. In: J. Bermudez and A. Millar (eds.): *Reason and Nature: Essays in the Theory of Rationality*. Oxford: Oxford University Press.

Castilho, M. A., O. Gasquet, and A. Herzig: 1999, 'Formalizing action and change in modal logic I: the frame problem'. *Journal of Logic and Computation* **9**(5), 701–35.

Cohen, P. R. and H. J. Levesque: 1990, 'Intention is choice with commitment'. *Artificial Intelligence* **42**, 213–261.

Danto, A.: 1965, 'Basic Actions'. *American Philosophical Quarterly* pp. 141–148.

Davidson, D.: 1980, 'Agency'. In: *Essays on Actions and Events*. New York: Oxford University Press.

Davis, L.: 1979, *Theory of Action*. Englewood Cliffs, N. J.: Prentice-Hall.

Dignum, F. and R. Conte: 1998, 'Intentional Agents and Goal Formation'. In: M. P. Singh, A. Rao, and M. Wooldridge (eds.): *Agent Theories, Architectures, and Languages: Agent Theories, Architectures, and Languages (ATAL-97)*. Berlin: Springer Verlag, pp. 231–243.

Dunin-Keplicz, B. and R. Verbrugge: 2002, 'Collective intentions'. *Fundamenta Informaticae* **51(3)**, 271–295.

Elgesem, D.: 1993, *Action Theory and Modal Logic*. PhD thesis, Dept. of Philosophy, University of Oslo: .

Elster, J.: 1979, *Ulysses and the sirens*. Cambridge: Cambridge University Press.

Fagin, R., J. Halpern, Y. Moses, and M. Vardi: 1995, *Reasoning about Knowledge*. Cambridge: MIT Press.

Frith, C., S. Blakemore, and D. Wolpert: 2000, 'Abnormalities in the awareness and control of action'. *Philosophical Transactions of the Royal Society of London: Biological Sciences* **355**, 1771–1788.

Gabbay, D., A. Pnueli, S. Shelah, and J. Stavi: 1980, 'On the temporal analysis of fairness'. In: *Proc. Seventh ACM Symposium on Principles of Programming Languages*. Las Vegas, NV, pp. 163 – 173.

Gerbrandy, J.: 1999, *Bisimulations on Planet Kripke*. The Netherlands: PhD thesis, University of Amsterdam.

Ginet, C.: 1990, *On Action*. Cambridge: Cambridge University Press.

Goldblatt, R.: 1992, *Logics of Time and Computation, 2nd edition*. Stanford, California: CSI Lecture Notes.

Goldman, A.: 1970, *A Theory of Human Action*. Englewood Cliffs NJ: Prentice-Hall.

Grosz, B. and S. Kraus: 1996, 'Collaborative Plans for Complex Group Action'. *Artificial Intelligence* **86**(2), 269–357.

Harel, D., D. Kozen, and J. Tiuryn: 2000, *Dynamic Logic*. Cambridge: MIT Press.

Harman, G.: 1986, *Change in View: principles of reasoning*. Cambridge: The MIT Press.

Herzig, A. and D. Longin: 2004, 'C&L intention revisited'. In: D. Dubois, C. Welty, and M.-A. Williams (eds.): *Proc. 9th Int. Conf. on Principles on Principles of Knowledge Representation and Reasoning(KR2004)*. pp. 527–535, AAAI Press.

Hornsby, J.: 1980, *Actions*. London: Routledge & Kegan Paul.

Horty, J. F. and N. Belnap: 1995, 'The deliberative stit: A study of action, omission, and obligation'. *Journal of Philosophical Logic* **24(6)**, 583–644.

Israel, D., J. Perry, and S. Tutiya: 1991, 'Actions and movements'. In: *Proc. 12th International Joint Conference on Artificial Intelligence (IJCAI'91)*. San Mateo, CA, pp. 1060–1065, Morgan Kaufmann.

Jordan, M. I. and D. M. Wolpert: 1999, 'Computational Motor Control'. In: M. Gazzaniga (ed.): *The Cognitive Neuroscience*. Cambridge: MIT Press.

Kanger, S.: 1971, 'New Foundations for Ethical Theory'. In: R. Hilpinen (ed.): *Deontic Logic: Introductory and Systematic Readings*. Dordrecht: D.Reidel, pp. 36–58.

Konolige, K. and M. E. Pollack: 1993, 'A representationalist theory of intention'. In: *Proc. 13th International Joint Conference on Artificial Intelligence (IJCAI 93)*. San Francisco, CA, pp. 390–395, Morgan Kaufmann.

Locke, J.: 1989, 'An Essay concerning Human Understanding'. Oxford: Clarendon Press.

McCann, H.: 1986, 'Rationality and the Range of Intention'. *Midwest Studies in Philosophy* **10**, 191–211.

McCann, H.: 1991, 'Settled Objectives and Rational Constraints'. *American Philosophical Quarterly* **28**, 25–36.

Mele, A. R.: 1992, *Springs of Action: Understanding Intentional Behavior*. Oxford: Oxford University Press.

Meyer, J. J. C., W. van der Hoek, and B. van Linder: 1999, 'A Logical Approach to the Dynamics of Commitments'. *Artificial Intelligence* **113(1-2)**, 1–40.

Miller, K. and G. Sandu: 1997, 'Weak Commitments'. In: G. Holmstron-Hintikka and R. Tuomela (eds.): *Contemporary Action Theory, vol.2: Social Action*. Dordrecht: Kluwer Academic Publishers.

O'Shaughnessy, B.: 1973, 'Trying (as the Mental "Pineal Gland")'. *Journal of Philosophy* **70**, 365–86.

Pacherie, E.: 2006, 'Towards a dynamic theory of Intentions'. In: S. Pockett, W. P. Banks, and S. Gallagher (eds.): *Does Consciousness causes Behavior? An investigation on the Nature of Volition*. Cambridge: MIT Press.

Panzarasa, P., N. Jennings, and T. J. Norman: 2002, 'Formalising collaborative decision making and practical reasoning in multi-agent systems'. *Journal of Logic and Computation* **12(1)**, 55–117.

Pörn, I.: 1977, *Action Theory and Social Science: Some Formal Models*. Dordrecht: Synthese Library 120, D. Reidel,.

Proust, J.: 2005, *La nature de la volonte'*. Paris: Folio-Gallimard.

Rao, A. S. and M. P. Georgeff: 1991a, 'Asymmetry thesis and side-effect problems in linear time and branching time intention logics'. In: *Proc. Twelfth International Joint Conference on Artificial Intelligence (IJCAI'91)*. pp. 498–504.

Rao, A. S. and M. P. Georgeff: 1991b, 'Modelling rational agents within a BDI-architecture'. In: *Proc. 2nd International Conference on Principles of Knowledge*

Representation and Reasoning (KR'91). San Mateo, CA, pp. 473–484, Morgan Kaufmann.

Reiter, R.: 2001, *Knowledge in action: logical foundations for specifying and implementing dynamical systems*. Cambridge: MIT Press.

Santos, F., J. Carmo, and A. Jones: 1997a, 'Action concepts for describing organised interaction'. In: *Thirtieth Annual Hawai International Conference on System Sciences*. Los Alamitos, California, pp. 373–382, IEEE Computer Society Press.

Santos, F., A. Jones, and J. Carmo: 1997b, 'Responsibility for Action in Organizations: a Formal Model'. In: G. Holmstron-Hintikka and R. Tuomela (eds.): *Contemporary Action Theory, vol.2: Social Action*. Dordrecht: Kluwer Academic Publishers.

Scherl, R. B. and H. Levesque: 2003, 'Knowledge, action, and the frame problem'. *Artificial Intelligence* **144**, 1–39.

Schroeder, S.: 2001, 'The concept of Trying'. *Philosophical Investigations* **24(3)**, 213–227.

Searle, J.: 1983, *Intentionality: An Essay in the Philosophy of Mind*. New York: Cambridge University Press.

Segerberg, K.: 1989, 'Bringing it about'. *Journal of Philosophical Logic* **18**, 327–347.

Segerberg, K.: 1992, 'Getting started: Beginnings in the Logic of Action'. *Studia Logica* **51(3-4)**, 347–378.

Sellars, W.: 1967, *Science and Metaphysics: Variations on Kantian Themes*. London: Routledge and Kegan Paul.

Shoham, Y.: 1993, 'Agent-oriented programming'. *Artificial Intelligence* **60**, 51–92.

Singh, M. and N. Asher: 1993, 'A logic of intentions and beliefs'. *Journal of Philosophical Logic* **22**, 513–544.

Thomason, R.: 2000, 'Desires and defaults: A framework for planning with inferred goals'. In: *Proc. Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR 2000)*. San Francisco, CA, pp. 702–713, Morgan Kaufmann.

Van Benthem, J. and E. Pacuit: 2006, 'The Tree of Knowledge in Action: Towards a Common Perspective'. In: G. Governatori, I. Hodkinson, and Y. Venema (eds.): *Proceedings of Advances in Modal Logic Volume 6 (AiML 2006)*. pp. 87–106, College Publications.

Van der Hoek, W., W. Jamroga, and M. Wooldridge: 2007, 'Towards a Theory of Intention Revision'. *Synthese* **155(2)**, 265–290.

Van Linder, B., van der Hoek, and J.-J. C. W., Meyer: 1998, 'Formalising abilities and opportunities'. *Fundamenta Informaticae* **34**, 53–101.

Von Wright, G. H.: 1963, *Norm and Action*. London: Routledge and Kegan.

Von Wright, G. H.: 1972, 'On so-called practical inference'. *The Philosophical Review* **15**, 39–53.

Wallace, R. J.: 2001, 'Normativity, Commitment, and Instrumental Reason'. *Philosophers' Imprint* **1(3)**, 55–117.

Wooldridge, M.: 2000, *Reasoning about rational agents*. Cambridge: MIT Press.