

To Attempt and To Try: for a cognitive theory of Action

Technical Report

Emiliano Lorini¹, Cristiano Castelfranchi²

¹University of Siena, Cognitive Science Doctorate
Via Roma 47, Siena, 53100 Italy
e.lorini@istc.cnr.it

²Institute of Cognitive Science and Technology–CNR
Viale Marx 15, Rome, 00137 Italy
c.castelfranchi@istc.cnr.it

This work is supported by the EU project “MindRACES: from reactive to anticipatory cognitive embodied systems”.

Introduction

Some symbolic AI models and logics, for example BDI models (Rao & Georgeff, 1992), are conceived as explicit and operational models of the intentional pursuit. If this is true then they should be able to take into account notions such as success, failure, prevention, expectation, attempt. We have a formal and computational model of intentional action and their related mental representations. This model should be able to give us a theory of Attempts as well as a theory of expectations.

Many philosophers have dealt with the notions of Attempts and Trying. Several views and approaches have been proposed. The main objective of this work is to provide an analysis of this notion which is crucial for a theory of Intentional Action. Especially the subjective anticipatory notion of Attempt will be investigated. This kind of Attempt will be distinguished from another kind of Attempt: the common “nucleus” of all actions (intentional and not intentional).

Several kinds of subjective Attempt will be identified and described. It will be shown that in the logical structure of every subjective Attempt a negative expectation is implied. Depending on the kind of Attempt taken into consideration the attempting agent has either a negative expectation with respect to the result (or state of the world) that he intends to achieve or with respect to the correctness of a certain action relative to an intended result, or with respect to feasibility of an action that has been planned for achieving an intended result.

Besides, it will be shown that under particular assumptions those negative expectations intrinsically connected to every attempt are generally conceivable as expectations about a possible failure or as expectations about a possible prevention. Especially the notion of Prevention will be disentangled and his causal structure will be analysed. It will be argued that for defining Prevention correctly, the identification of an event that disables the conditions for the execution of the action and for the success of the action is needed.

But before the discussion of the previous action-centered concepts, a clarification of the notion of intentional action, “intending to do something” and “intending that something happens” will be provided. A special emphasis will be assigned to develop a causal grammar based on the distinction between Movements of the body and Delegated external events. The grammar will be used for the identification of Basic and Complex Actions, Strong and Weak Actions. Moreover, a recursive analysis of Actions and Vehicle Actions will be provided.

Finally on the basis of the analysis of the notion of intentional action a specific part of the work will be devoted to clarify the notion of “Intention that”.

1. Kinds of attempt: a preliminary account

The purpose of the present work is to propose a theory of the mental and extra-mental notion of *Attempt* which has been introduced in philosophy of action. We will mainly focus in this analysis on the second notion of Attempt (Attempt β).

But first of all let us distinguish Attempt β which is the topic of our analysis from *Attempt α* .

Attempt α

The present category *Attempt α* has the same status of the notion of Trying defined by O'Shaughnessy (1972) and by Hornsby (1980). O'Shaughnessy for instance identifies the two following laws for describing Trying.

Trying as mental 'pineal gland' (The Psychophysical Law).

Trying, in being essentially a cause of physical phenomenon and a linchpin of consciousness, serves a crucial bridge function between mind and body...

The Psycho-Psycho Law

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.

Attempt α is the command of execution of the action in the physical world, the starting command for the execution of the action. In our view *Attempt α* is the intermediate element of the causal process which leads from the present-directed intention (Bratman, 1987) to the execution of the action in the external world.

Attempt β

Attempt β is the "dubious" initiation of an action aimed at achieving a certain intended result (or state of the world). We specify three kinds of Attempt β depending on the object of the doubt:

- The object of the doubt is the achievement of the intended result;
- The object of the doubt is the correctness of the action with respect to the intended result;
- The object of the doubt is the feasibility of the action.

We will show that Attempts β can either fail or succeed.

In other words Attempt β is the initiation of an action aimed at achieving a certain result with a negative expectation¹ concerning either:

- the achievement of the result;
- the correctness of the action with respect to the intended result;
- the execution of the action (aimed at achieving the intended result).

¹ In our previous works (Miceli & Castelfranchi 2002, Castelfranchi & Lorini 2003) we have defined Positive and Negative Expectations (Fears, Hopes etc...) as complex mental states (we did not introduce Expectations as an additional primitive). Expectations were built on former ingredients (beliefs and goals/intentions) in order to have mental states that preserve both properties, epistemic and conative. We have argued that Expectations have a specific

2. The Basic model of Action

The “problem of description” is relevant whenever we try to identify what a certain agent is doing. The “problem of description” concerns the 1,..., n interpretations that we can have whenever we try to identify the action executed by a certain agent x. Given two agents in the world, an executer and an observer, the action of the executer can be described differently by the executer himself or by the observer. So which is the ontological status of the Action? The “problem of description” has been raised by Anscombe (1956). An intention can be described either by the “owner” of the intention agent x or by an external observer agent y. The observer (agent y) can offer an interpretation of the action executed by agent x and given the interpretation of agent x’s action agent y identifies agent x’s intention. The description given by agent y could be completely different from the description given by agent x.

In the present analysis we are assuming that the **action** can never be separated from its own **result**. We assume that always an action is done by an agent for achieving a certain result and so the result of the action is intrinsically connected with the action (for instance if the agent intends to “open the window” then “the window is open” is the result of the action) and without the identification of the result of the action neither the action can be identified.

The four main assumptions of the present analysis are the following.

1. the identification of the **action** done by the agent depends on the identification of an achieved **result of the action**;
2. the identification of the **action** done by the agent depends on the attribution of an **intention to do the action** to the agent;
3. the identification of an achieved **result of the action** depends on the attribution of an **intention that the result of the action holds** to the agent;
4. the attribution of an **intention to do the action** to the agent depends on the attribution of an **intention that the result of the action holds** to the agent;



Figure 1

Similarly with Von Wright (1971) we assume here that the action is identified by what the agent **INTENDS TO DO** and the result (state of the world) that the agent **INTENDS THAT IT HOLDS** by means of the

functional role in practical reasoning that is better understood by defining those mental states in a compositional fashion.

action. Differently from Von Wright we assume that an action can be conceived as a complex of bodily movements and delegated external event² so the notion of **INTENTION TO DO** that we propose here it is composed by an **Intention To Do** some bodily movement and the **Intention That** some (delegated) external event **Happens**.

The problem of the identification of the action is considered here as a *nomination problem*, i.e. in order to say what the agent x has effectively done by means of the action I need to identify the intended result that the agent has achieved by means of the action and that the agent intended to achieve by means of the action.. An action can be named if and only if it is connected with an achieved result for which the action is done.

In the following part of the paragraph we will:

- 1) propose a general typology of Actions and Vehicle Actions through an analysis of the different kinds of causal relations between their components.
- 2) propose an analysis of the notion of “Intention that”.

2.1 The ingredients for an intentional theory of Action

The following are the general ingredients for an intentional theory of Action.

- The **Causal antecedents** of the intended result of the action can be decomposed as follows:
 - **Controlled movements of the body**: bodily movements³ that the agent does intentionally⁴.
 - **Delegated external events**: external events (either natural events or other agent’s actions) that happen and that the agent “intends that” happen.

Controlled movements of the body and delegated external events can be grouped together in order to form:

- **Vehicle Action**: causal sequence of n (with $n > 0$) controlled bodily movements that the agent does (intentionally) in order to achieve an intended result and that is part of a more extended causal sequence of $n + m$ (with $m > 0$) controlled bodily movements and/or delegated external events that the agent does (intentionally) in order to achieve an intended result.
- **Specific Action**: causal sequence of n (with $n > 0$) controlled bodily movements and/or delegated external events that the agent does (intentionally) in order to achieve an intended result and that is not part of

² According to Von Wright the only causal antecedents of the result of the action are bodily movements.

³ A more general label would be (instead of “controlled bodily movements”) “controlled emission of kinetic energy in the environment”.

⁴ The notion of “Intending to do something” (the “Pure Intending”) and the notion of “Doing something intentionally” have been clearly distinguished in Davidson (1980). We adopt here the Simple View in which we assume that “an agent does something intentionally if and only if “the agent intends to do that”. The Simple View has been strongly criticized by Bratman (1987). According to Bratman an agent does intentionally also what he believes will be a consequence of his intended action, moreover he does intentionally all spontaneous actions that are done during the execution of the intended action. This position is slightly different from the position presented of Searle (1983). According to Searle an agent can have the “Intention in Action to do a certain action” without the need of a previous “Prior Intention to do that action” (all intentional actions have intentions in action but not all intentional actions have prior intentions). This seems to apply to automatic and spontaneous *micro-actions* that were not planned by the agent (they were not object of Prior Intentions) but that are executed during the execution of the *macro-action* that was the object of the Prior Intention (Bratman would say that those spontaneous micro-actions are done intentionally but they are not object of Present-directed Intentions).

a more extended causal sequence of $n + m$ (with $m > 0$) controlled bodily movements and/or delegated external events that the agent does (intentionally) in order to achieve an intended result.

Here the **Result of the Action** is: the state of the world that the agent intends that holds.

On the side of the intention we adopt the distinction given by Bratman (1987) between Present-directed and Future-directed Intentions. Indeed in order to establish what the agent has done it is necessary to have a notion of present-directed intention that can be attributed to the agent. According to Bratman a Present-directed Intention is the intention to do something *now* whereas a Future-directed Intention is the intention to do something *later*⁵.

That causal and compositional analysis should be developed both on the side of the objective action and on the side of the INTENTION TO DO the action.

We have now all the tools for re-establishing in a better way the main assumptions that drive the present analysis. The new assumptions should be considered as a deeper specification of the four assumptions given above.

1. The identification of the **action** depends on the identification of an **“actional” causal complex** of controlled movements of the body and external events.
2. The identification of an **“actional” causal complex** of controlled movements of the body and external events depends on the identification of an achieved **result of the action**.
3. The identification of an achieved **result of the action** depends on the attribution of a **future-directed intention that the result of the action holds**.
4. The identification of an **“actional” causal complex** of controlled movements of the body and external events depends on the attribution of a **present-directed intention** to do the first element (s) of the “actional” causal complex relative to⁶ the **Future-directed intention** (to do or that) about the second element of the causal complex,..., relative to the **Future-directed intention** (to do or that) about the last element of the causal complex⁷.
5. The identification of the **action** depends on the attribution of a **present-directed intention to do the action**.
6. The attribution of a **present-directed intention** to do the first element (s) of the “actional” causal complex relative to the **Future-directed intention** (to do or that) about the second element of the causal complex,..., relative to the **Future-directed intention** (to do or that) about the last element of the causal complex depends on the attribution of a **Future-directed intention that the result of the action holds**.

⁵ Bratman assigns to the mental state of future-directed intention the feature of *persistence* (if an agent intends to do an action later then the agent is committed to do that action and he will give up his intention if and only if particular conditions hold, i.e. given particular conditions for reconsidering the intention).

⁶ The notion of relativized intention as specified in Cohen & Levesque (1990) is used to express the mean-end relation in the decomposition of the plan (action a is intended by the agent relative to an intended result p i.e. the agent is committed to do action a and if he discover that he does not intend that p then he drops the commitment with the action a).

⁷ We have included relativized intentions in order to take into account what Searle (1980) calls “planned regularity” and that is introduced in order to avoid that an intentional action is identified whenever unplanned causal chains take part in the achievement of the final intended result and planned causal chains do not take part in the achievement of the final intended result.

Notice that in the present analysis the meaning of “Doing a Plan (or an Action)” does not require all the conditions identified by Pollack (1990) for providing the meaning of the notion of “Having a Plan (or an Action)”. According to Pollack the six conditions for “Having a Plan” are the following. An agent A “Has a plan to do B”, that consists in doing some set of acts α , provided that: 1) A believes that he can execute each act in α ; 2) A believes that executing the acts in α will entail the performance of B; 3) A believes that each act in α plays a (causal) role in his plan; 4) A intends to execute each act in α ; 5) A intends to execute α as a way of doing B; 6) A intends each act in α to play a (causal) role in his plan. In the present analysis the doxastic conditions provided by Pollack are not included. Indeed we argue here that those conditions are necessary for understanding why a certain agent has a specific intention relative to the intention to achieve the final result but are not necessary to specify what the agent is actually doing (or has done). Finally the relativized intention allows to model the causal role of each component of the plan with respect to the other components of the plan and with respect to the result of the action (it merges the last Pollack’s conditions 4 6).

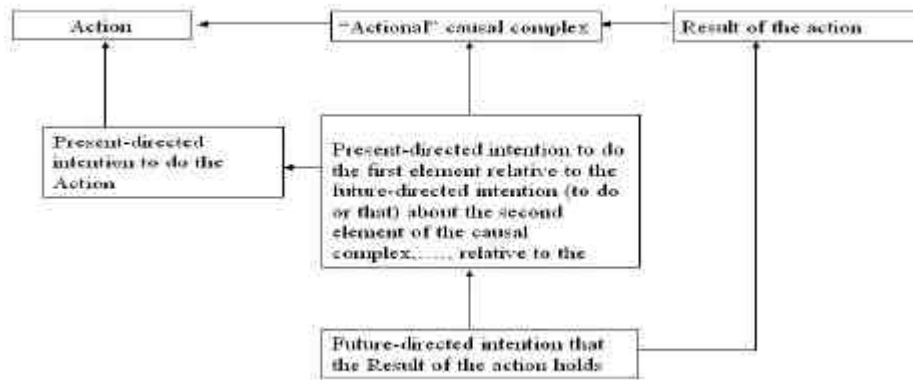


Figure 2

2.2 The Typology of the actions⁸

In the present paragraph we want to propose an analysis of the notion of Action and Plan. We will consider two dimensions for the analysis: the dimension Basic versus Complex and the dimension Strong versus Weak. The first dimension concerns the type and the number of elementary events that compose a given Action, the second dimension concerns the type of causal relations between the elements that compose the Action. In the following part the distinction between Specific Action and Vehicle Action is not developed, we merely develop a typology for a general notion of **Action**:

a causal sequence of n (with $n > 0$) controlled bodily movements and/or delegated external events that the agent does (intentionally) in order to achieve an intended result.

Just consider the fact that by means of the following analysis it is easy to double the typology in order to identify for each class of Action the related notions of Specific Action and Vehicle Action (see above).

Moreover we want to remark that the seven assumptions given in the previous paragraph are the theoretical bases of the following analysis.

Definition 2.1: Basic Action a with respect to a state of the world p (BA a p).

We can say that action a is a Weak Basic Action with respect to the state of the world p that agent x has done (**Weak-BA x a p**) if and only if a is a dynamic sequence of n elements where *one and only one element* is a controlled movement of the body and *all other elements* are external events. Moreover the n elements happens as a **causal complex d** organised according to a **criteria of weak causality (instrumentality) with respect to the state pupper bounded by a controlled movement of the body** (WC-Causal Complex d). Moreover, the WC-causal complex d determines the state of the world p and agent x had before that the causal complex d happens, a general “INTENTION TO DO” the WC-Causal Complex d relative to the “INTENTION THAT” p HOLDS” where the “INTENTION TO DO” the WC-Causal Complex d is represented as a sequence of relativized intentions (to do or that) with respect to the elements of d that reproduces the causal relations of d itself (i.e. if an element i of the causal complex is a cause for another element j of the causal complex then there is an intention (to do or that) about element i relative to an intention (to do or that) about element j).

We can say that action a is an agent's Strong Basic Action with respect to the state of the world p that agent x has done (**Strong-BA x a p**) if and only if a is a dynamic sequence of n elements where *one and only one element* is a controlled movement of the body and *all other elements* are external events. Moreover the n elements happens as a **causal complex d** organised according to a **criteria of strict causality (instrumentality) with respect to the state pupper bounded by a controlled movement of the body** (SC-Causal Complex d). Moreover the SC-causal complex d determines the state of the world p and agent x had before that the causal complex d happens, a general “INTENTION TO DO” the SC-Causal Complex d relative to the “INTENTION THAT” p HOLDS” where the “INTENTION TO DO” the SC-Causal Complex d is represented as a sequence of relativized intentions (to do or that) with respect of the elements of d that respect the causal relations of d itself.

⁸ Notice that the distinction Basic versus Complex in following Typology is similar to Davidson's (1980) distinction between Basic Actions and Complex Actions. However in the analysis of Davidson delegated external events as components of actions are not taken into account.

Definition 2.2: Complex Action a with respect to a state of the world p (CA a p).

We can say that action a is a Weak Complex Action with respect to the state of the world p that agent x has done (**Weak-CA x a p**) if and only if a is a dynamic sequence of n elements where *at least two elements* are controlled movements of the body and *all other elements* are external events. Moreover the n elements happens as a **causal complex d** organised according to a **criteria of weak causality (instrumentality) with respect to the state pupper bounded by a controlled movement of the body** (WC-Causal Complex d). Moreover, the WC-causal complex d determines the state of the world p and agent x had before that the causal complex d happens, a general “INTENTION TO DO” the WC-Causal Complex d relative to the “INTENTION THAT” p HOLDS” where the “INTENTION TO DO” the WC-Causal Complex d is represented as a sequence of relativized intentions (to do or that) with respect of the elements of d that respect the causal relations of d itself.

We can say that action a is an agent's Strong Complex Action with respect to the state of the world p that agent x has done (**Strong-CA x a p**) if and only if a is a dynamic sequence of n elements where *at least two elements* are controlled movements of the body and *all other elements* are external events. Moreover the n elements happens as a **causal complex d** organised according to a **criteria of strict causality (instrumentality) with respect to the state pupper bounded by a controlled movement of the body** (SC-Causal Complex d). Moreover the SC-causal complex d determines the state of the world p and agent x had before that the causal complex d happens, a general “INTENTION TO DO” the SC-Causal Complex d relative to the “INTENTION THAT” p HOLDS” where the “INTENTION TO DO” the SC-Causal Complex d is represented as a sequence of relativized intentions (to do or that) with respect of the elements of d that respect the causal relations of d itself.

Definition 2.3a: The two criteria of causality

The n elements in a set $S = \{1, \dots, n\}$ are organized according to the **criteria of weak causality (instrumentality) with respect to state pupper bounded by a controlled movement of the body** if and only if each element in S is an indirect cause for p . Moreover, for each element in S if that element is an external event then it is always indirectly caused by another element of S that is a controlled movement of the body and for each element in S if that element is a controlled movement of the body then it is directly caused by the agent.

The n elements in a set $S = \{1, \dots, n\}$ are organized according to the **criteria of strict causality (instrumentality) with respect to state pupper bounded by a controlled movement of the body** if and only if there is only one element $i \in S$ that is a direct cause for p and there is only element k in S such that k is a direct cause of only another element in S , and k is a controlled movement of the body and k is not (directly or indirectly) caused by any element of S but it is directly caused by the agent and finally for any other element $i \in S' = \{\{1, \dots, n\} - \{i\}\} - \{k\}$ if i is an external event then i is a direct cause of only another element of $S' = \{\{1, \dots, n\} - \{k\}\}$ and is directly caused by only one element of $S'' = \{\{1, \dots, n\} -$

$\{i\}$ and if i is controlled movement of the body then i is a direct cause of only another element of $S' = \{\{1, \dots, n\} - \{k\}\}$, i is directly caused by only one element of $S'' = \{\{1, \dots, n\} - \{i\}\}$ and i is directly caused by the agent.

It is relevant to notice that the notion of causal relation and criteria of (weak or strict) causality previously expressed can be expressed in terms of necessity relation and/or in terms of sufficiency relation.

Indeed in order to express the fact that a certain component of a plan is causally related with another component of the same plan, at least three different kinds of causal relations could be taken into account:

- The fact that an element a of the causal complex d happens is a necessary and sufficient condition for the fact that another element β of d happens (NS).
- The fact that an element a of the causal complex d happens is a necessary and but not sufficient condition for the fact that another element β of d happens (NnotS).
- The fact that an element a of the causal complex d happens is a sufficient but not necessary condition for the fact that another element β of d happens (SnotN).

We assume here that in order to identify an action as a causal complex d is enough that the causal relations between the elements of d are expressed in one of the previous ways (NS, NnotS, SnotS) or in simpler forms (N, S). In the following definitions some more restrictions are imposed. The meaning of the restrictions is explained afterwards and it depends on the way an agent comes to intend (to or that) and to have relativised intentions on the basis of his beliefs about instrumentality and beliefs about Ability.

Definition 2.3b: The two criteria of causality revisited in terms of Sufficiency and Necessity Condition

The n elements in a set $S = \{1, \dots, n\}$ are organized according to the **criteria of weak causality (instrumentality) with respect to state pupper bounded by a controlled movement of the body** if and only if each element in S is an indirect cause for p and for each element $i \in S$ if that element is an external event then there is always at least another element $j \in S$ such that j is a controlled movement of the body and that is an indirect necessary condition for i . For each element $i \in S$ if that element is an external event then all elements of S directly causing it are necessary condition for i . For each element $i \in S$ if i is a controlled movement of the body then all the (eventual) elements of S plus the agency are (taken together) compatible conditions of i and all elements of S directly causing i are necessary conditions for i .

The n elements in a set $S = \{1, \dots, n\}$ are organized according to the **criteria of strict causality (instrumentality) with respect to state pupper bounded by a controlled movement of the body** if and only if there is only one element $i \in S$ that directly causes p and that element is a direct necessary condition for p , and there is only element $k \in S$ such that k is not directly caused and the agency is a compatible condition for k and k directly causes only another element $l \in S$, and if l is an external event then k is a

necessary condition for it otherwise k is a necessary condition for it and k and the agency are (taken together) a compatible condition for it.

For any other element $v \in S'' = \{\{1, \dots, n\} - \{i\}\} - \{k\}$ if v is an external event then v directly causes only another element $q \in S' = \{\{1, \dots, n\} - \{k\}\}$ and v is a necessary for q and v is together with the agency a compatible condition for q ; and v is directly caused by only one element $s \in S''' = \{\{1, \dots, n\} - \{i\}\}$ and s is a necessary condition for v .

For any other element $v \in S'' = \{\{\{1, \dots, n\} - \{i\}\} - \{k\}\}$ if v is a controlled movement of the body then v directly causes only another element $q \in S' = \{\{1, \dots, n\} - \{k\}\}$ and v is a necessary condition for q and v is together with the agency a compatible condition for q ; and v is directly caused by only one element $s \in S''' = \{\{1, \dots, n\} - \{i\}\}$ and s is a necessary condition for v and s and the agency are (taken together) a compatible condition for v .

For summarising, Weak Causality is used for representing every kind of causal complex of controlled movements of the body and external events on the basis of the following constraints.

- 1) Every controlled movement of the body $m1$ is caused by some controlled movements of the body and/or some external events where the controlled movements and/or external events are taken separately necessary conditions for $m1$ and the agency and/or those controlled movements of the body and/or external events (directly causing it) are taken together compatible conditions for $m1$.
- 2) Every external event $e1$ is caused either by some controlled movements of the body $m1$ and/or some external events where the controlled movements and/or external events are taken separately necessary conditions for $m1$.
- 3) The result of the action is caused either by some controlled movements of the body $m1$ and/or some external events where the controlled movements and/or external events are taken separately necessary conditions for the result.

On the other side Strict Causality is used for representing a sequential causal chain (special kind of causal complex) of controlled movements of the body and external events on the basis of the following more stringent constraints.

- 1) Every controlled movement of the body $m1$ is caused by only one controlled movement of the body $m2$ or only one external event $e1$ where $m2$ or $e1$ is a necessary condition for it and the agency and $m2$ or $e1$ are (taken together) a compatible condition for $m1$.
- 2) Every external event $e1$ is caused either by only one controlled movement of the body $m1$ or only one external event $e2$ where the controlled movement $m1$ or external event $e2$ is a necessary condition for $m1$.
- 3) It exists a first element of the causal sequence that is movement of the body and that is not directly caused and the agency is a compatible condition for it.
- 4) It exists a last element of the causal sequence that is either a movement of the body or an external event and that directly causes the result of the action p and that is a necessary condition for p .

5) The result of the action is caused either by only one controlled movement of the body m_1 or only one external event e_2 where the controlled movement m_1 or external event e_2 is a necessary condition for the result.

The idea of organizing the criteria of weak and strict causality in terms of the Necessity Condition with respect to the Causation that goes from controlled movements of the body to controlled movements of the body, from external events to external events, from controlled movements of the body to external events, from external events to controlled movements of the body, from controlled movements of the body to the result of the action and from external events to the result of the action, derives from the way we conceive the practical syllogism. We assume here the validity of the practical schema proposed by Von Wright (1978) who argues that the essence of Aristotelian practical syllogism is best captured by the construction

Agent i intends to make it true that a

Agent i believes that, unless it does action β , it will not achieve this

Therefore Agent i intends to do action β .

The practical syllogism applies in our view to the process that brings from an Intention that an external event happens or from an Intention that the result of the action holds or from an Intention that a movement of the body is realized to the instantiation (through selection) of a candidate (relativized) Intention that.

We are accepting here a Rule that expresses the fact that in order to have a candidate relativized “Intention that” the agent has to evaluate on the basis of a Necessity criteria.

On the other side the idea of organizing the criteria of weak and strict causality in terms of the Compatibility Condition with respect to the Causation that goes from the agency and the controlled movements of the body to the controlled movements of the body, from the Agency and external events to the controlled movements of the body derives from the way we conceive the way we pass from a candidate Intention that a movement of the body is realized to the Intention to execute that movement of the body.

We are assuming here that a necessary condition (the constraint) for having an Intention to execute a movement of the body is the fact of not having the belief about the Complete incapability of executing that movement (see the model of Castelfranchi, 1996 for an analysis of the role of beliefs about Ability and Incapability in the practical reasoning⁹).

As we will show later (in paragraph 3.1.1) the Ability to do a movement of the body m_1 is substantially defined either

- as a notion of Mediated Ability requiring a notion of Causal Mediator that it is actually holding in the world and a Sufficiency Condition such as the following “always if I have the present-directed intention to execute the movement m_1 and the causal Mediator holds then I execute successfully movement m_1 ”;

⁹ Our theoretical position is very close to the Asymmetry Thesis (Bratman, 1987 Rao & Georgeff, 1991). According to the Asymmetry thesis in fact it is irrational for an agent to Intend to do something that he believes to be completely incapable to do whereas it is acceptable from a rational point of view that an agent Intend to do something and he does not believe to be fully able to do it. The Asymmetry thesis is again applicable for understanding how we pass from a candidate Intention that an external event happens to an Intention that an external event happens. It simply requires that

- or as a not Mediated notion (Not Mediated ability) that is simply defined in terms of a Sufficiency Condition “always if I have the present-directed intention to execute the movement m1 then I execute successfully movement m1”.

Now, we argue that the notion of Complete Incapability to execute a movement of the body is defined either:

- as a notion of Mediated Complete Incapability requiring a notion of Causal Negative Mediator that is actually holding in the world and a Negative Sufficiency Condition such as “always if I have the present directed intention to execute the movement m1 and the causal Mediator holds then I fail to execute movement m1”;

- or as a not Mediated Notion (Not Mediated Complete Incapability) that is defined again in terms of a Negative Sufficiency Condition “always if I have the present-directed intention to execute the movement m1 then I fail to execute successfully movement m1”.

The requirement that in order to Intend to execute a controlled movement of the body m1 it must hold the fact that the agent does not believe to be completely incapable (either in Not Mediated or in Mediated sense) of executing the controlled movement m1 implies that in order to identify an action on the basis of the identification of an Present-Directed Intention to execute a controlled movement of the body m1 (relativized to some other Intention -to or that- about some other elements of the causal complex that identifies the action) it must hold the fact that the Present-directed Intention to execute the controlled movement of the body m1 (that corresponds with the Agency-related notion of “*Attempt a* to execute m1” as defined in chapter 1) + the eventual Causal Mediator are Compatible with respect to the controlled movement m1. The causal notion of Compatibility differently from the stricter causal notions of Sufficiency and Necessity simply requires that the Present-directed Intention to execute the movement of the body m1 + the eventual Causal Mediator have a weak causal relation with the movement of the body m1. Two kinds of weak causal relations are conceivable.

- Mediated Compatibility: “it exists a Causal Mediator that is actually holding in the world and “sometimes if I have the present-directed intention to execute the controlled movement m1 and the causal Mediator holds then I succeed in executing the controlled movement m1”.

- Not Mediated Compatibility: “sometimes if I have the present-directed intention to execute the controlled movement m1 then I execute successfully the controlled movement m1”¹⁰.

the agent believes that event to be possible (that it is equivalent to the acceptance of the fact that the agent does not believe that event will necessarily happen and the refusal that the agent believes that event to be impossible).

¹⁰

As we will show next the theory of Bringing About is good enough to model the previous kinds of strict and weak causal relations that must exist in the causal complex of movements of the body and external events where the causal complex is a necessary conceptual construct for the identification of the Action.

The logic of “Bringing about” we refer here is the one proposed by Porn (1977) and extended in Santos et al. (1997). We do not review the semantic of the formal language. We will just discuss the meaning of some operators and we will show how the logic of “Bringing about” allows to model the causal relations in terms of necessary/sufficient conditions/compatible conditions.

Let us assume that At is a set of *movement of body symbols* and Eev is a set of *external events symbols*.

The operator E_x is the “bringing about” causal operator that can be defined both in terms of “Necessity for something that an agent does” (Positive Condition) and in terms of “Counteraction Conditionality”(Negative Condition). This logic is based on the following modal operators.

$D_x m_1$ means that “it is necessary for something which x does that m_1 ”,

$D'_x m_1$ means that “but for the agent x 's action it would have been the case that m_1 ”

$C_x m_1$ means that “it is compatible with everything the agent x does that m_1 ”

$C'_x m_1$ means that “for x 's activity it might not be the case that m_1 ”.

$E_x m_1$ means that “agent x brings about the movement of the body m_1 ” that is equivalent to

$D_x m_1 \dot{\cup} C'_x \neg m_1$ whereas

$F_x m_1$ means that “agent x lets it be the case that m_1 ” that is equivalent to

$C_x m_1 \dot{\cup} C'_x \neg m_1$ that is equivalent to $m_1 \dot{\cup} C'_x m_1$ (the notion of consequence of action)

and finally

$Nec_x m_1$ means that “it is a practical necessity for the agent x that m_1 ” that is equivalent to

$D_x m_1 \dot{\cup} D'_x m_1$.

Porn introduces in his theory of the Automaton the operator of “bringing about something by bringing about something else” as well as the operator of “letting it be the case that by bringing it about something”.

$E_x (m_1, e_1)$ means that “agent x brings about e_1 by bringing about m_1 ” that is equivalent to

$\exists A \exists s \in S_A (e_1 = O(s, m_1) \dot{\cup} E_x m_1 \dot{\cup} D_x s \dot{\cup} C'_x e_1)$ where A is the automaton, S_A is the set of the possible internal states of the Automaton, O is the output function of the automaton given an internal state and an input.

$F_x (m_1, e_1)$ means that “agent x lets it be the case that e_1 by bringing about m_1 ” that is equivalent to

$\exists A \exists s \in S_A (e_1 = O(s, m_1) \dot{\cup} E_x m_1 \dot{\cup} s \dot{\cup} C'_x e_1)$ where A is the automaton, S_A is the set of the possible internal states of the Automaton, O is the output function of the automaton given an internal state and an input.

Necessity causal relations and/or sufficiency causal relations can be identified and modelled by means of the Bringing about's logic at two different levels :

1. Between the Agency and a movement of the body (or an external event);

2. Between a movement of the body and an external event (or an external event and a movement of the body, or between two external events, or between two movements of the body).

In order express the different kinds of causal relations between the agency and a movement of the body the different primitive operators of the Logic can be used as shown in the following table.

<p>Necessity of the Agency x with respect to an event e1</p> $\neg C'_x e_1 = D'_x \neg e_1$
<p>Sufficiency of the Agency with respect to an event e1 = Incompatibility of the Agency with respect to an event $\neg e_1$</p> $D_x e_1$
<p>Not Necessity of the Agency with respect to an event e1</p> $C'_x e_1$
<p>Not Sufficiency of the Agency with respect to an event e1 = Compatibility of the Agency with respect to an event $\neg e_1$</p> $C_x \neg e_1 = \neg D_x e_1$

On the other side in order to express the different kinds of causal relations between a movement of the body and an external event (or an external event and a movement of the body, or between two external events, or between two movements of the body) we need to consider more carefully the notion of Transmission of Agency and Automaton provided by Porn.

An Automaton can be defined as a function $y = O(s, x)$ where x (INPUT) and y (OUTPUT) are either movements of the body or external events or states of the world and $s \in S_A$ is an internal state of the Automaton. Let us define in the following table Necessity relations and Sufficiency relations.

<p>Necessity of the event e1 with respect to the event e2</p> <p>Necess (e1, e2)</p> <p>It does not exist a $s' \in S_A$ and $x \in Eev$ or $x \in At$ such that $x ? e_1$ and $e_2 = O(s', x)$</p>
<p>Sufficiency of the event e1 with respect to the event e2</p> <p>Suff (e1, e2) = Incompatibility of event e1 with respect to an event $\neg e_2$</p> <p>For each $s \in S_A$ $e_2 = O(s, e_1)$ for all $s \in S_A$</p>
<p>Not Necessity of the event e1 with respect to the event e2</p> <p>Not-Necess (e1, e2)</p> <p>It does exist a $s' \in S_A$ and $x \in Eev$ or $x \in At$ such that $x ? e_1$ and $e_2 = O(s', x)$</p>
<p>Not Sufficiency of the event e1 with respect to the event e2</p>

Not-Suff (e1, e2) = Compatibility of event e1 with respect to an event -e2

It exists a $s' \in S_A$ and $x \in Eev$ or $x \in At$ such that
 $x = O(s', e1) \text{ AND } x \neq e2$

Given the previous characterisation of different kinds of causal relations we are able to provide several kinds of complex expressions such as

$(E_x(m_1, e_1) \dot{\cup} \text{Suff}(e_1, e_2) \dot{\cup} \text{Necess}(e_1, e_2))$ or $(F_x(m_1, e_1) \dot{\cup} \text{Suff}(e_1, e_2) \dot{\cup} \text{Necess}(e_1, e_2))$ that gives information both about the causal relations between the Agency and the element m_1 , between the Agency and the element e_1 and about the causal relations between the element m_1 and the element e_1 .

Complex recursive expressions such as

$(E_x(x_1, \dots, x_n) \dot{\cup} \text{Suff}(x_1, x_2) \dot{\cup} \text{Necess}(x_1, x_2) \dot{\cup} \dots \dot{\cup} \text{Necess}(x_{n-1}, x_n))$

that means “agent x brings about x_n by bringing about x_{n-1} and agent x brings about x_{n-1} by bringing about x_{n-2} ...and agent x brings about x_2 by bringing about x_1 and x_1 is a necessary and sufficient condition for x_1, \dots , and x_{n-1} is a necessary condition for x_n ” or such as

$(F_x(x_1, \dots, x_n) \dot{\cup} \text{Suff}(x_1, x_2) \dot{\cup} \text{Necess}(x_1, x_2) \dot{\cup} \dots \dot{\cup} \text{Necess}(x_{n-1}, x_n))$

that means “agent x lets it to be that x_n by letting to be that x_{n-1} and agent x lets to be that x_{n-1} by letting to be that x_{n-2} ...and agent x lets to be that x_2 by bringing about x_1 and x_1 is a necessary and sufficient condition for x_1, \dots , and x_{n-1} is a necessary condition for x_n ”

could also be built on the bases of the present analysis.

It is clear now that the Bringing about formalism is powerful enough to characterize the causation of the agency with respect to the terminal elements of the causal complex that identifies an action as well as the causal relations (expressed in terms of necessary conditions) between the elements of the causal complex (controlled movements of the body and external events).

Indeed as shown above the operator E_x is already able to capture the sufficiency condition of the agency with respect to the movement of the body (that operator is also enriched by the further counteraction condition that according to our view expresses the Not Necessity of the Agency with respect to the negation of event or of the movement of the body to which the operator E_x is applied). Moreover, in the “Bringing about” theory it is possible to specify the Compatibility of the agency with respect to the movement of the body m_1 (by an opportune use of the operator C_x).

The only problem is that the Bringing About formalism does not provide tools for modelling the intentional side of the action. This is the main problem of the theory: the absence of a modelling of important notions such as future-directed and present-directed intention, relativized intention.

A different possible way to model causations between the agency and the terminal elements of the causal complex that identifies the action (on the basis of the previous arguments causation is specified here in terms

of the Compatibility condition) and between the elements of the causal complex (on the basis of the previous arguments causation is specified here in terms of the necessary condition) is the following.

A certain agent x is a compatible condition for a given movement of the body m1 if and only always if the agent has the present-directed intention to execute m1 (if the Agent Attempt a to execute the controlled movement of the body m1 then m1 is executed next).

A certain movements of the body m1 (or external event e1) is a necessary condition for a given movement of the body m2 (or an external event e2) if and only if always if m1 is not executed (or event e1 does not happen) then m2 is not executed (or external event e2 does not happen).

A dynamic and temporal approach very close to the one proposed by Van der Hoek & Wooldridge (2003) and Van Linder et al. (1998) empowered by means of a detailed logic of Intentions could be used to express those kinds of causal relations. That approach could also be translated by defining counterfactuals relations and dependences in the sense of Lewis (1973).

2.3 Some examples for the Typology

Figure 3 below presents four possible causal structures of Weak Basic Action, Strong Basic Action, Weak Complex Action and Strong Complex Action. For each causal structure a concrete example is given.

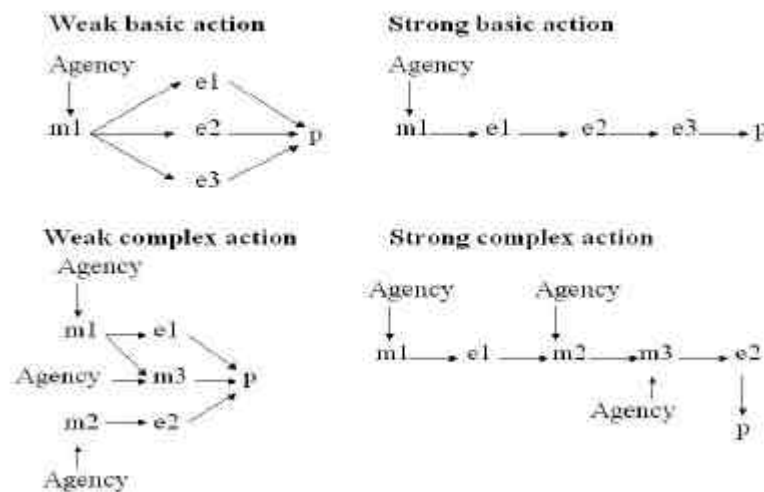


Figure 3

Weak Complex Action

Imagine that an agent x whistles (m_1) in order that agent y realizes to be called (e_2). At the same time agent x raises the right arm (m_2) in order to move the hand in a higher position of the space (m_3) and in order that agent y sees him (e_1). Here $m_1 \rightarrow e_1$, $m_1 \rightarrow m_3$, $m_2 \rightarrow e_2$, $e_1 \rightarrow p$, $m_3 \rightarrow p$ = "agent y believes that agent x is calling at him".

Weak Basic Action

Imagine agent x is inside the room and whistles (m_1) in order that agent y , agent z , agent m enter into the room (e_1, e_2, e_3). External event e_1, e_2, e_3 together determine that p = "the group of friends is gathered inside the room". Here $m_1 \rightarrow e_1$, $m_1 \rightarrow e_2$, $m_1 \rightarrow e_3$, $e_1 \rightarrow p$, $e_2 \rightarrow p$ and $e_3 \rightarrow p$.

Strong Basic Action

Imagine agent x is inside the room and whistles (m_1) in order that agent y enters into the room (e_1) and in order that agent z sees that agent y enters into the room (e_2) and finally in order that agent z too enters into the room (e_3). Here $m_1 \rightarrow e_1 \rightarrow e_2 \rightarrow e_3 \rightarrow p$ where p = "the group of friends is gathered inside the room".

Strong Complex Action

Imagine agent x is stuck inside the room 1 and whistles (m_1) in order that agent y (who has the key of room 1) open the door of the room 1 (e_1) and in order to go into room 2 (m_2) and in order to switch on the hitting (m_3) and finally in order that the thermostat increases the temperature of the room (e_2) and finally in order to get p = "the temperature of the room is warm". Here $m_1 \rightarrow e_1 \rightarrow m_2 \rightarrow m_3 \rightarrow e_2 \rightarrow p$.

2.4 Some intermediate results

Let us assume here the following transitive property for causal relations.

Axiom of transitivity

For each element $a_1, a_2, a_3 \in \text{Eve U At}$

If (a_1 CAUSES a_2) AND (a_2 CAUSES a_3) THEN
(a_1 CAUSES a_3)

Theorem*

i) If Strong-BA a p then Weak-BA a p ; ii) If Strong-CA a p then Weak-CA a p .

Proof: I will prove here only part i) of the theorem.

Through the consecutive application of the transitivity property it follows that

$(m_1 \text{ CAUSES } e_2 \text{ CAUSES } \dots \text{ CAUSES } e_n \text{ CAUSES } e_{n+1} \text{ CAUSES } p) \rightarrow$

$(m_1 \text{ CAUSES } (e_2 \wedge \dots \wedge e_n \wedge e_{n+1}) \text{ CAUSES } p)$.

The following FIGURE 4 resumes the relation between previous Strict-notions and Weak Notions.

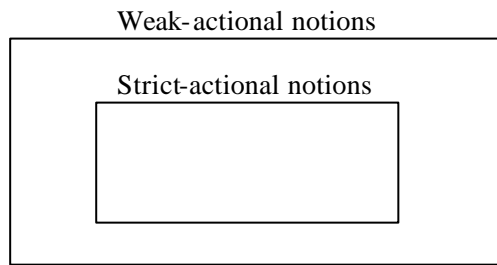


Figure 4

2.5 The notion of “Intention that”

As we have already shown in the previous paragraphs the “INTENTION TO DO” an action has an “internal” component and an “external” one: an “Intention to do” some controlled bodily movements and an “Intention that” some external events or some other agents’ actions will happen. For instance the INTENTION TO DO the action “fill the sink of water ” is composed by the “Intention to” execute the movement “turn the tap” and the “Intention that” the external event “the water falls in the sink” will happen. The INTENTION TO DO the action “fill the sink of water ” is relative to the “INTENTION THAT” “the sink will be full of water” HOLDS.

We want to discuss here in a more extensive way the notion of “Intention that”.

The first issue: *The three kinds of “Intention that”.*

As in Sellars (1967) and as it follows from our previous analysis we can characterize three types of Intention That¹¹:

- An **INTENTION THAT** a *Result of the action* (state of the world) **HOLDS**;
- An **Intention That** a *Natural Event* **Happens**;
- An **Intention That** an *Agent Action* **Happens**.

The second issue: “Intention that” and self-realizing results

When an agent has an INTENTION THAT a certain result HOLDS then either the agent has planned an action for causing the object of the INTENTION THAT so he INTENDS TO DO a specific action relative to the INTENTION THAT the result HOLDS, or he INTENDS THAT it exists an action such that it happens and it is a necessary condition for the object of the INTENTION THAT¹² and by adopting again the Asimmetry Thesis (Bratman, 1987), in order that the agent INTENDS THAT it exists an action such that it happens and it is a necessary condition for the object of the INTENTION THAT then it is required that the

¹¹ In Grosz & Kraus (1995) for instance Intention That is only with respect to other agent actions.

¹² See again Sellars’ distinction (1966) between Intentions *to do* and Intentions *that something be the case*: “*It shall be the case that-p* has the sense, when made explicit, of *I shall do that which is necessary to make it the case that-p*”.

agent DOES NOT BELIEVES THAT it does not exist an action such that it happens and it is a necessary condition for the intended result.

There are cases in which the previous doxastic requirement can not be satisfied: all those cases in which the agent conceives the result as a self-realising, i.e. the agent believes that it does not exist any action such that it is a necessary condition for the intended result. In all those cases the agent can not INTEND THAT it exists an action such that it happens and it is a necessary condition for the result and from this it follows that the agent can not INTEND THAT the result HOLDS.

Self-realising results are results whose achievements can not be controlled by the agent. From a subjective (doxastic) point of view wished (or wanted) results that the agent believes to be self-realising can not be the objects of an INTENTION THAT.

Let us consider more accurately the case in which the agent INTENDS THAT a certain result HOLDS, he does not intend to do any specific action but he INTENDS THAT it exists an action such that it happens and it is a necessary condition for the result. Let us imagine that the agent comes to believe that the result is self-realising. The agent can not INTENDS anymore THAT the result HOLDS. We have a “regression” of the INTENTION THAT the result HOLDS to the level of the GOAL THAT the result HOLDS. The agent can now merely delegate (and effectively delegates) to external events the full achievement of the intended result (this is called *Weak Delegation* in Falcone & Castelfranchi, 1998) since he believes that those external events will cause the achievement of the final intended result and believes that those external events will happen by themselves. The agent is still motivated to find an action for achieving the wished (wanted) final result but at the moment he cannot INTEND THAT it exists an action such that it happens and it is a necessary condition for the result since he believes that that action does not exist. Also in this case we have a “regression” of the INTENTION THAT it exists an action such that it happens and it is a necessary condition for the result to the level of the GOAL THAT it exists an action such that it happens and it is a necessary condition for the result.

Whenever the agent comes to believe that his own contribution is needed in order to achieve the final result, the agent will be able to have again the regressed INTENTIONS THAT and he will be able to plan an action for bringing about those external events that are still according to him determinants of the final (again intended) result (this is called *Mild Delegation* in Falcone & Castelfranchi, 1998).

The further process is also quite relevant. The situation in which the agent has planned and INTENDS TO DO a specific action and he discovers that the intended bodily components that compose the planned action are useless since the intended result will be realised by itself, by means of a sequence of external events. In this case the agent had an INTENTION TO DO a specific Action relative to an INTENTION THAT a certain result HOLDS; afterwards the agent has simply a GOAL THAT a sequence of Natural Events happens and the GOAL THAT a certain result HOLDS.

Imagine for instance the following situation. Agent x INTENDS TO water the lawn in the garden in order THAT the grass will be wet. At the end of the first stage agent x has the INTENTION TO water the lawn in the garden where the intention can be decomposed as follows: “Intention to” ask to the gardener to water the lawn \rightarrow “Intention that” the gardener waters the garden \rightarrow INTENTION THAT “the lawn is wet” HOLDS..

Before that agent x asks to the gardener, agent x sees that the gardener is re-directing the water pipe towards the lawn. Agent x believes at the second stage that he does not need to make a request to the gardener. Agent x reconsiders the “Intention to” ask to the gardener and after the reconsideration he has not more the INTENTION TO water the lawn in the garden. Both his intentions have regressed to the level of goals and the agent has now a GOAL THAT “the gardener waters the garden” → GOAL THAT “the grass is wet” HOLDS.

From the previous arguments and from the arguments provided in the previous chapter it also follows that an agent can INTEND THAT a certain result HOLDS also in those cases in which he believes that the planned action is not a sufficient reliable condition for the achievement of the result or in those cases in which he believes that there is not an action that is a sufficient reliable condition for the achievement of the result. The only requirement for INTENDING THAT a certain result HOLDS is that the agent believes that the planned action is a necessary condition for the achievement of the result or in those cases in which he believes that there is an action that is necessary reliable condition for the achievement of the result (as we have seen this excludes all those cases in which the result is conceived as a self-realising). According to our view an agent who is playing a game with a fair dice can perfectly INTENDS THAT “a six is thrown”. Indeed the agent is perfectly able to conceive an action = movement $m1$ = “raise the arm”; movement $m2$ = “throws the dice”; event $e1$ = “the dice rolls” that according to his beliefs is a necessary condition for the result = “a six is thrown”¹³.

The third issue: *The separation of the Intentions.*

Let me consider with more attention the following causal relations.

Movement-body 1 CAUSES External Event 1

External event 1 CAUSES final intended result p

External Event 2 CAUSES Movement-body 1

What can we say? Which is the structure of the associated intentions?

According to the definitions of complex Action given above and the restriction due to the criteria of strong instrumentality upper bounded by a movement of the body, we can not identify in the previous causal structure a Strong Complex Action. The only thing we can do is to separate the INTENTION THAT the final result p HOLDS from the INTENTION TO DO the Strong Basic Action composed by MovementBody 1 and External Event 1 and this INTENTION TO DO from the “Intention that” External Event 2 happens.

Again we have on one side an “Intention that” External Event 1 happens inside the more general INTENTION TO DO the Strong Basic Action and that “Intention that” is a strong one; on the other side we

¹³ Our position is here slightly different from the position of Mele & Moser (1994) who argue that a necessary condition for saying that the agent *does A Intentionally* [in our view: for saying that the agent INTENDS that some result p HOLDS] is the fact that the agent believes that the plan selected for doing A is a *suitably reliable mean* for doing A [in our view: the fact that the agent believes that the plan selected for achieving p is a *suitably reliable mean* for achieving p], where the notion of *suitably reliable mean* is related with the notion of Sufficiency. According to our theoretical proposal the requirement of Sufficiency proposed by Mele & Moser is too strong.

have an “Intention that” the External Event 2 happens that is not inside a more general INTENTION TO DO a certain action and so it has a weak one.

2.6 The distinction Micro versus Macro

Till now I have used the notion of Controlled Movement of the body and External Event as the minimal units of intentional actions. We have shown that an agent must have a repertoire of those minimal units that form compositionally Basic Actions and Complex Actions. Here we want to make a further distinction. We call *the minimal components that the agent can use to form (by planning and delegating) either Basic Actions or Complex Actions in advance with respect to their execution* **Macro-movements** and **Macro-events**.

We call

the minimal components that the agent use in doing either Basic Actions or Complex Actions either **Micro-movements** and **Micro-events**.

This distinction is very close to the one of Searle (1980). We agree with the theoretical position that: many movements of the body that the agent executes (does) intentionally and many external events that the agents delegates intentionally (i.e. that are object of Intentions in Action) were not object of previous future-directed intentions. The problem is to find a model of action that can integrate the Macro-level with the Micro-level. We leave this crucial issue to further elaboration of the theory.

We want only to argue here that micro-events and micro-actions that are con-causes with the planned action (the macro-action) for the achievement of a certain intended result p always exist. Somehow, an agent always *relies on* at a mere motor-level on some process of the external world for achieving the intended result p and the same agent *executes* unplanned movements of the body during the execution of the planned action. Imagine for instance the case the agent was standing on the floor of a room and intended to open the door of the room at t1. The agent started an action at t1 for opening the door. It reasonable to assume that natural causal antecedents such as “the wind blows at 100 km/h” or “the door is swinging due to the presence of external factors” always exists. Imagine now a car driver that has the previous intention of “making a curve” and during the execution of the action changes gear. According to Searle (1980) the driver is intentionally changing the gear even he did not have the previous intention to do it. The same line reasoning applies to events that are not Delegated external events in the future directed intention but on which the agent relies in the phase of motor coordination during the execution of the action. For instance the car driver who has to stop in order to avoid the dog in the middle of the street, will intentionally brake and will swerve by relying either on some movement of the dog or on the absence of movements of the dog.

The decomposition of actions in controlled movements of the body and delegated external events is merely a decomposition at the intentional level. An action can be further decomposed. Indeed, every delegated external event can be decomposed in micro-events on which the agent relies on and every controlled movement of the body can be decomposed in micro-movements that the agent executes in doing that controlled movement.

The important point here is provided by the following statement that predicates about the interactive reality of the agent with his environment.

Every execution of an intentional action A implies some form of coordination with the external world, i.e every execution of an intentional action A implies at least a reliance on not intended external events.

3. Prevention and Failure

In this part of the analysis we want to provide a clarification of the notion of Prevention and Failure. Those categories are in fact strongly related with the anticipatory-based notion of Attempt β that this work is aimed at clarifying. Indeed the negative expectations logically implied by every attempt are generally conceivable as expectations about a possible failure and under particular assumptions (that will be clarified) the negative expectations logically implied by every attempt are generally conceivable as expectations about a possible prevention. We will also provide a clarification of the notions of Ability and Opportunity and we will show their relations with the notions of Prevention, Conditions for Action Execution and Condition for the Success of the Action. A

Let us extract first of all the following general valid schema which shows three necessary and sufficient condition for doing a certain action a.

SCHEMA a

If

Condition A

The agent has the future directed intention to do action a as a mean for achieving intended result p AND

Condition B1

The agent will have the future directed intention to do action a as a mean for achieving intended result p till the time the action is sent to execution

AND

Condition C

It will not prevented that the agent will achieve the intended result p by doing the intended action a (selected for achieving p).

then

the agent will achieve p by doing action a

Condition B1 in previous predictive **SCHEMA a** is very important. Indeed we can never state that the three mental states expressed in Condition A are sufficient condition for “starting” a certain action a unless we assume that the three mental states has hold till the time the action is sent to execution.

We want to discuss in the following paragraphs the notion of prevention that we have introduced in condition c of SCHEMA a.

We want first to stress a crucial issue. A very general meaning of Prevention (Impedimento) as regards the intended action a to be done at t1 of agent x for achieving the intended result p at t2 should be composed by two parts:

- Whatever external event e1 which causes that the agent who has a future directed intention to do a certain action a for achieving the intended result p either gives up the intention to do a for achieving p (and so chooses another mean) or gives up the intention to achieve p (so he gives up the intention about the mean (a) for achieving p) so that event e1 causes that the Action is not responsible for achieving p.
- Whatever external events e1 which causes that the intended action does not realize or that the action does not bring about the intended result.

By taking into account this double general definition for Prevention we could translate the predictive SCHEMA a in the following general SCHEMA β.

SCHEMA β

If

Condition A

The agent has the future directed intention to do action a as a mean for achieving intended result p AND

Condition B

It will not be prevented that the agent will achieve the intended result p by doing the intended action a (selected for achieving p).

then

the agent will achieve p by doing action a

The previous general definition of prevention includes the fact that the agent can give up his own intention to do a. There could be many different external events which determine that the agent gives up his own future directed intention to do a (persuasion by other agents, further evaluations about the action by the agent etc...).

In the following analysis we will only focus on the second kind of prevention (and preventing). So we will be merely interested in taking into account the Prevention with respect to the execution of the action and we will neglect the Prevention with respect to the Reconsideration of the Intention.

Before going to examine deeply the notion of Prevention let us conclude with the following statement:

Every time it will be prevented that the agent will achieve p by doing a then the action has to be re-described.

How the action of the agent needs to be re-described? What does the agent really do? We follow here the idea of von Wright (1971): *what the agent really does is limited to what the agent does intending to do the action.*

3.1 A definition of Prevention

We will provide in this paragraph a detailed analysis of the notion of Prevention. We will use a dynamic and temporal approach extended to cover Intentional notions.

Let us start first from the first type of prevention, let us first of all explain what we mean with the sentence “**Condition ? prevented that the agent achieved the intended result p by doing the action a selected to achieve it**”. We can at most identify the following two cases.

<p>Definition 3.1</p> <p>Definition for “It has been prevented that the agent x achieved the intended result p by doing the action a selected to achieve it, by preventing the execution of the planned action a” (Prevention of the planned achievement of the Intended result p by preventing the execution of the planned action a).</p> <p>For each agent x, state of the world p, action a</p> <p>Achiev-PREVENT.BYActExec-PREVENT e1 x a p if and only if</p> <p>The agent had a present-directed intention to do action a^{14} now relative to the future directed intention to achieve p after the execution of a.</p> <p>Moreover,</p> <p>It exist conditions a and β and ? such that</p> <p>- Condition ? held and condition β held;</p> <ol style="list-style-type: none"> 1. Always if the agent has the present-directed intention to do action a and condition a hold, action a happen next (strong sufficiency modality). 2. Always if action a happens and condition β hold, p hold after the execution of a (strong sufficiency modality). <p>This first two conditions imply the following more compact condition.</p> <p>1+2. Always if the agent has the present-directed intention to do action a, condition a and condition β</p>	<p>Definition 3.2</p> <p>Definition for “It has been prevented that the agent x achieved the intended result p by doing the action a selected to achieve it, by preventing the success of the action a” (Prevention of the planned achievement of the Intended result p by preventing the success of the planned action a).</p> <p>For each agent x, state of the world p, action a</p> <p>Achiev-PREVENT.BYActSuc-PREVENT e1 x a p if and only if</p> <p>The agent had a present-directed intention to do action a now relative to the future directed intention to achieve p after the execution of a.</p> <p>Moreover,</p> <p>It exist conditions a and β and ? such that</p> <p>- Condition ? held and condition a held;</p> <ol style="list-style-type: none"> 1. Always if the agent has the present-directed intention to do action a and condition a hold, action a happen next (strong sufficiency modality). 2. Always if action a happens and condition β hold, p hold after the execution of a (strong sufficiency modality). <p>This first two conditions imply the following more compact condition.</p> <p>1+2. Always if the agent has the present-directed intention to do action a, condition a and condition β hold, p holds after the execution of a.</p>
---	--

¹⁴ As specified in the previous chapter the intended action a should be understood in terms of a causal complex of movements of the body and external events.

hold, p holds after the execution of a.	3. Always if the agent has the present-directed intention to do action a but condition a does not hold then action a does not happen next (Necessity Modality).
3. Always if the agent has the present-directed intention to do action a but condition a does not hold then action a does not happen next (Necessity modality).	4. If the agent does not have the present-directed intention to do action a but condition a holds then it might be the case that action a does not happen next (Counter-action modality).
4. If the agent does not have the present-directed intention to do action a but condition a holds then it might be the case that action a does not happen next (Counter-Action Modality).	5. Always if action a happens next but condition β does not hold then p does not hold after the execution of a.
5. Always if action a happens next but condition β does not hold then p does not hold after the execution of a.	6. If action a does not happen next but condition β holds then it might be the case that p does not hold after the execution of a.
6. If action a does not happen next but condition β holds then it might be the case that p does not hold after the execution of a.	7. Always if Condition ? holds, condition β does not hold.
7. Always if Condition ? holds, condition a does not hold.	8. Always if Condition ? did not hold, condition β holds.
8. Always if Condition ? does not hold, condition a holds.	

The previous conditions a and β and ? can be defined respectively as it follows.

Condition for action execution. It is the condition that guarantees the causal (or counterfactual) dependence of the action with respect to the present directed intention, i.e. that is together with the Present-directed Intention (again Attemp a) a Sufficient Condition for the execution of the Action and that is Necessary condition for the execution of the action.

As we have shown in chapter 2 an intended action is identifiable in terms of a causal complex of controlled movements of the body and delegated external events. This implies that there are two kinds of Conditions for Action execution: conditions for the execution of the intended movement of the body (i.e. the fact that the agent is not tied), conditions for the realization of the delegated external event.

Condition for the success of the action. It is the condition (not intended by the agent) that guarantees the the causal (or counterfactual) dependence of the result of the action with respect to the execution of the action, i.e. that is together with the Action a Sufficient Condition for the achievement of the result of the action and that is a Necessary condition for the achievement of the result of the action.

Preventing condition. It is the condition on which the negation of condition a or the negation of condition β counterfactually depends.

Condition a and β and ? can be represented as disjunctions of conjunctions $(a_1 \wedge \dots \wedge a_n) \vee \dots \vee (a_m \wedge \dots \wedge a_v)$ where each element a_i is either a state of the world or the fact that a specific event happens.

As shown in the previous chapter an intended plan is specified as a sequence of intended (to do) movements of the body and intended (that) external events. From this it follows that the general condition for action execution is decomposable as a set of conditions for the execution of a single movement of the body in the sequence that identifies the intended action and conditions for the realization of a single external event in the same sequence. Each condition for the execution (or the realization) of an element of the sequence that identifies is conceivable a logical disjunctions of conjunctions.

$(a_1 \wedge \dots \wedge a_s) \vee \dots \vee (a_m \wedge \dots \wedge a_v)$ CONDITION OF EXECUTION OF
THE FIRST ELEMENT OF THE SEQUENCE

.....

$(?_1 \wedge \dots \wedge ?_n) \vee \dots \vee (?_r \wedge \dots \wedge ?_j)$ CONDITION OF EXECUTION OF
THE LAST ELEMENT OF THE SEQUENCE

The previous definition given for CASE1 contains a specific definition for **Planned Action Prevention**.

Definition 3.3

Definition for “It has been prevented that the agent executed the action a planned for achieving the intended result p” (Prevention of the execution of the planned action a).

For each agent x , state of the world p , action a

ActExec-PREVENT x a p if and only if

The agent had a present-directed intention to do action a now relative to the future directed intention to achieve p after the execution of a .

Moreover,

It exist conditions a and $?$ such that

- Condition $?$ held;

1. Always if the agent has the present-directed intention to do action a and condition a hold, action a happen next (strong sufficiency modality).
2. Always if the agent has the present-directed intention to do action a but condition a does not hold then action a does not happen next (Necessity Modality).
3. If the agent does not have the present-directed intention to do action a but condition a holds then it might be the case that action a does not happen next (Counter-action Modality).
4. Always if **Condition $?$** holds, condition a does not hold.
5. Always if **Condition $?$ does not** hold, condition a holds.

Moreover, CASE 1 “**Prevention of the planned achievement of the Intended result p by preventing the execution of the planned action a**” covers two interesting cases.

Definition 3.1.a

Prevention of the planned achievement of the Intended result p by preventing the execution of the planned action a without Prevention of the achievement of the Intended result p.

CASE 3.1.b

Prevention of the planned achievement of the Intended result p by preventing the execution of the planned action a with Prevention of the achievement of the Intended result p.

On the other side CASE 2 always implies that the intended result p is not achieved (this is self-contained in the definition).

Consider for example the following conversation between two persons that clarifies previous CASE 1B (**TO BE TRANSLATED**).

A *Perché la finestra è aperta? L'agente ha aperto la finestra?*

B *L'agente non ha aperto la finestra, la finestra si è aperta da sé (è stato il vento ad aprirla). Anche se l'agente non avesse eseguito la sequenza di movimenti "di aprire la finestra" che tu credi essere responsabile dell'apertura della finestra, la finestra si sarebbe aperta in ogni caso ma d'altra parte se la finestra non si fosse aperta da sé l'agente non avrebbe potuto aprirla dato che l'agente era impedito, ma se non fosse stato impedito avrebbe aperto sicuramente la finestra.*

A *Per certi aspetti all'agente è stato impedito di aprirla?*

B *Direi piuttosto che all'agente è stato impedito di essere l'effettivo fautore del risultato dell'azione.*

A good example for CASE 1A is the following soccer scenario. The player intends to score and relative to it he plans to hit the ball in a certain way and to direct the ball towards the goal. Unfortunately defender 1 touches the ball and changes its trajectory (i.e. defender 1 prevents that the trajectory of the ball is the planned-anticipated one, defender 1 prevents that the causal complex intended by the player happens) but defender 2 touches the ball and changes again the trajectory of the ball, re-directing it towards the goal and guaranteeing that the player scores. However in this scenario a con-causal relation can be identified between the partial intended action that effectively happens and defender 2's intervention (event). The partial intended action and defender 2's intervention taken together are conceivable as event e3 that appears in condition 7 of the previous definition.

We should investigate all those cases of **'Prevention of the achievement of the Intended result p'** without neither **"Prevention of the planned achievement of the Intended result p by preventing the execution of the planned action a"** (defined in CASE 1) nor **"Prevention of the planned achievement of the Intended result p by preventing the success of the planned action a"** (defined in CASE 2).

However all those cases are left apart in this paper. We want only to address a common general structure.

Definition 3.4

Definition for "It has been prevented that the agent achieved the intended result p" (Prevention of the achievement of the Intended result p).

For each agent x, state of the world p, action a

AchievePREVENT x a p if and only if

The agent had a present-directed intention to do action a now relative to the future directed intention to

achieve p after the execution of a .

Moreover,

It exist conditions d and $?$ such that

- **Condition ? held**

1. Always if condition d holds, result p holds after the execution of a .
2. Always if condition d holds, result p holds after the execution of a .
3. Always if **Condition ? does not** hold, condition d does not hold.
4. Always if **Condition ? does not** hold, condition d holds.

3.1.1. Ability as a Sufficient condition for action execution

Let us to specify more carefully the differences between the previous notion of Conditions for action execution and Conditions for the success of the action and the notions of Ability. We argue here it exist a strong similarity between those notions and so they should be conceptually analysed. The important points we want to stress here are: 1) the fact that different kinds of Abilities should be identified depending on the degree of Control that the agent actually has on the performance of the action or on the achievement of the intended result; 2) the fact that the Ability to execute a certain controlled movement of the body or the Ability to do a certain action are not necessary conditions but are special kinds of sufficient conditions for the execution of the movement of the body or for the execution of the Action or for the success of the Action. With respect to the second point, several authors have provided examples to deny that ability is a necessary condition for action. Kenny (1975) for example, argues that we cannot attribute the ability to hit the bull's eye to a hopeless dart-player who has hit only once in his career. Elgesem (1997) presents the case of a child who is learning to eat by himself. Suppose he succeeds at some point during the meal and that the mother is watching him to make sure that he gets fed. If the child does not succeed in feeding himself, the mother is ready to put food into his mouth anyway. In the example, there is not relevant possible alternative where it is not true that the child puts food in his mouth. Formal solutions to prevent that Ability to do an Action is a necessary for doing the action have been provided⁵.

We want to provide next a specific way to address the notion of Ability (and the notion of Opportunity) that does not incur into the problem of necessity condition risen by Kenny but that address the notion of Ability (and Opportunity) in terms of a Sufficient Condition for the execution of the Action (or in terms of Sufficient Condition for the success of the action).

Definition 3.5

Ability to execute a certain movement of the body. A certain agent x is able to execute a certain movement of the body m_1 if and only if

Mediated Ability

It exist a **mediating condition** ? such that **condition** ? holds and

1a. always if the agent has the present-directed intention to execute m1 and ? holds then movement m1 is executed next.

The following conditions specify a special kind of Mediated Ability in which it is required that the mediating condition holds but in which it exist a Sufficiency Relation between the present-directed intention and the mediating condition.

It exist a **mediating condition** ? such that

1b. always if the agent has the present-directed intention to execute m1 and ? holds then movement m1 is executed next.

2b. Always if the agent has the present-directed intention to execute m1 then ? holds.

Not mediated Ability

1. always if the agent has the present-directed intention to execute m1 then movement m1 is executed next;

It does not exist a **mediating condition** ? such that

2. always if the agent has the present-directed intention to execute m1 then ? holds;

3. always if ? holds then movement m1 is executed next.

Mediated Ability is formalized in order to consider the cases of causal mediation and parasite ability: all those cases in which the present-directed intention to execute m1 “exploits” some mediator (either some other movement of the body or some external events) for determining the m1’s execution¹⁶. Mediated abilities to do movements of the body are abilities that are not completely under the control of the agent whereas not mediates abilities to do movements of the body are abilities that are completely under the control of the agent¹⁷. Conditions 2 and 3 in the definition of Not mediated ability are given in order to exclude from the definitions all those cases of Mediated Sufficiency of the Present-directed Intention (special case 1b + 2b of Mediated Ability). Consider again the example provided by Elgesem of the baby who is not able to put the food in the mouth (movement m1) whose mother is helping him to feed. Possible situations in which the baby has the intention to put food in the mouth and in which the mother re-direct the movement of the body in the correct way whenever she recognises the intention in the mind of his baby (the mother executes some movement that determines the correct movement m1 of the child) are conceivable. We say that the baby is weakly able to put the food in the mouth. Indeed in all possible situations the mother (influenced by the intention of the child) is responsible for the execution of the movement.

¹⁵ See for instance the solutions proposed within the Seeing to it that (STIT) approach (Horty & Belnap 1995, Belnap 1991, Belnap & Perloff 1988).

¹⁶ The notion of Mediated Ability is completely omitted in Demolombe (2004) who merely focus on a special kind of Not Mediated Ability such that an agent is able to bring about something if and only if if the agent attempts to bring about it then it obtains.

¹⁷ For a discussion concerning the relation between abilities and control see again Elgesem (1994).

Definition 3.6

Ability to do a certain action identifiable as a causal complex of movements of the body and external events. A certain agent x is able to do a certain action identifiable as a constrained¹⁸ causal complex d of agency, movements of the body and external events if and only if

Mediated Ability¹⁹

(INFORMAL DEFINITION) The complete conjunction a of elements of the causal complex d that are direct necessary conditions and/or compatible conditions for another element e_1 of the causal complex and a **mediating condition** γ (that can be represented as a collection of external events that are not part of d) are together sufficient conditions for e_1 .

(FORMAL DEFINITION) For each controlled movement of the body m_1 in the causal complex d and for each logical conjunction a of elements (movements of the body and external events) of the causal complex d if (always if a does not hold then m_1 is not executed and it does not exist another element e_z of the causal complex d such that always if e_z does not hold then m_1 is not executed) then (it exists a mediating condition γ represented as a collection of external events that are not part of the causal complex d such that γ holds and always if the agent has the present-directed intention to execute m_1 , a holds, and the mediating condition γ holds then m_1 is executed after d). For each controlled external event e_1 in the causal complex d and for each logical conjunction a of elements (movements of the body and external events) of the causal complex d if (always if a does not hold then m_1 is not executed and it does not exist another element e_z of the causal complex d such that always if e_z does not hold then e_1 is not executed) then (it exists a mediating condition γ represented as a collection of external events that are not part of the causal complex d such that γ holds and always if a holds, and the mediating condition γ holds then e_1 happens after d).

Not mediated Ability

(INFORMAL DEFINITION) The complete conjunction a of elements of the causal complex that are direct necessary conditions and/or compatible conditions for another element e_1 of the causal complex are taken together sufficient conditions for that element.

(FORMAL DEFINITION) For each controlled movement of the body m_1 in the causal complex d and for each logical conjunction a of elements (movements of the body and external events) of the causal complex d if (always if a does not hold then m_1 is not executed and it does not exist another element e_z of the causal complex d such that always if e_z does not hold then m_1 is not executed) then (always if the agent has the

¹⁸ See Chapter 2 for the constraints on the kinds of Causations between the elements of the causal complex defining an Action.

¹⁹ The distinction between Mediated Ability and Not Mediated Ability to do an Action is very close to the under-specified distinction of Mele (2003) between Simple and Intentional Ability. According to Mele although an agent could be simply able to roll a six with a single toss of fair die, he can not be able to do that intentionally. We have argued (criticizing the position of Mele & Moser, 1994) in paragraph 2.5 that an agent could in principle intend to roll a six with a single toss of a fair die. Here we agree with Mele. Indeed we argue that an agent can not have the Not Mediated Ability to roll a six since it does not exist any conceivable sub-components of a causal complex that defines the action "rolls six with a single toss of fair die" such that it is sufficient condition for the last component of the causal complex that defines the action, that is the external event "it rolls a six". However, the agent could have in principle the Mediated Ability provided by the existence of a Mediating Condition such as "the wind blows in a certain way..." (notice that this Mediating Condition can not be intended and so it cannot be part of the causal complex that defines the action since the notion of fair die implies that the agent must conceive that event as a self-realizing one).

present-directed intention to execute m_1 and a holds then m_1 is executed after d). For each controlled external event e_1 in the causal complex d and for each logical conjunction a of elements (movements of the body and external events) of the causal complex d if (always if a does not hold then m_1 is not executed and it does not exist another element $e_{1,z}$ of the causal complex d such that always if $e_{1,z}$ does not hold then e_1 is not executed) then (always if a holds then e_1 happens after d).

Definition 3.7

The Opportunity to achieve a certain result p . A certain agent x has the opportunity to achieve a certain result p if and only

Mediated Ability

It exist a **mediating condition** $?$ and an action A that the agent is able to do either in a not mediated way or in a mediated way such that $?$ holds and always if action A is executed next and $?$ holds then p holds.

Not Mediated Opportunity

It exist an Action A that the agent is able to do either in a not mediated way or in a mediated way such that always if action A is executed next then p holds.

Definition 3.8

The Opportunity to achieve a certain result p by means of a specific action A . A certain agent x has the opportunity to achieve a certain result p by means of a specific action A if and only

Mediated Ability

the agent is able to do A either in a not mediated way or in a mediated way action and it exist a **mediating condition** $?$ such that $?$ holds and always if action A is executed next and $?$ holds then p holds.

Not Mediated Opportunity

the agent is able to do A either in a not mediated way or in a mediated way action always if action A is executed next then p holds.

3.2 Failure and Success

We will present next the definition of Failure in the general framework for an intentional theory of action. The notion of Failure is less specific than the notion Prevention: it does not need an accurate causal explanation. Three different kinds of Failure can be identified.

Definition 3.9

Failure to achieve an Intended result p .

For each agent x , state of the world p

the agent x has failed to achieve an intended result p if and only if

1. It exists an action a_1 such that agent x had the present directed intention to do action a_1 relative to the future directed intention to achieve the intended result p after the execution of a_1 ;

2. result p was not achieved after the execution of a1.

Definition 3.10

Failure to execute an intended action a1.

For each agent x, action a1

The agent x has failed to execute the action a1 planned for achieving an intended result p if and only if

1. It exists a result p such that agent x had the present directed intention to do action a1 relative to the future directed intention to achieve the intended result p after the execution of a1;
2. action a1 was not executed next.

Finally let us provide a definition for **Failure in the planned achievement of the Intended result p**. We can identify three sub-species (A, B, C).

Definition 3.11.a

Failure in the planned achievement of the Intended result p by failing to execute the planned action a1 with Success of the achievement of the Intended result.

For each agent x, state of the world p, action a1

The agent x has failed to achieve the intended result p by doing the action a1 selected to achieve it, by failing to executed the planned action a1 and he has succeeded to achieve the intended result if and only if

1. agent x had the present directed intention to do action a1 relative to the future directed intention to achieve the intended result p after the execution of a1 and p was achieved;
2. action a1 was not executed next and p was achieved.

Definition 3.11.b

Failure in the planned achievement of the Intended result p by failing to execute the planned action a with Failure in the achievement of the Intended result.

For each agent x, state of the world p, action a1

The agent x has failed to achieve the intended result p by doing the action a1 selected to achieve it, by failing to executed the planned action a1 and he has failed to achieve the intended result if and only if

1. agent x had the present directed intention to do action a1 relative to the future directed intention to achieve the intended result p after the execution of a1 and p was not achieved;
2. action a1 was not executed next and p was not achieved.

Definition 3.11.c

Failure in the planned achievement of the Intended result p by failing to achieve it.

For each agent x, state of the world p, action a1

The agent x has failed to achieve the intended result p by doing the action a1 selected to achieve it p by failing to achieve it if and only if

1. Agent x had the present directed intention to do action a1 relative to the future directed intention to achieve the intended result p after the execution of a1 and p was achieved;
2. action a1 was executed next and p was not achieved.

It is easy to prove that every **Prevention of the achievement of the Intended result p (def. 3.4)** implies a **Failure to achieve an Intended result p (def. 3.9)**, to prove that every **Prevention of the execution of the planned action a (def. 3.3)** implies a **Failure to execute an action a (3.10)**, every **Prevention of the planned achievement of the Intended result p by preventing the execution of the planned action a without Prevention of the achievement of the Intended result p (def. 3.1a)** implies a **Failure in the planned achievement of the Intended result p by failing to execute the planned action with Success of the achievement of the Intended result p (def. 3.11a)**, every **Prevention of the planned achievement of the Intended result p by preventing the execution of the planned action a with Prevention of the achievement of the Intended result p (def. 3.1b)** implies a **Failure in the planned achievement of the Intended result p by failing to execute the planned action with Failure in the achievement of the Intended result p (def. 3.11b)**, every **Prevention of the planned achievement of the Intended result p by preventing the success of the planned action a (def. 3.2)** implies a **Failure in the planned achievement of the Intended result p by failing to achieve the intended result (def. 3.11c)**. It is much more difficult to prove the opposite statements. Those statements can be proved only by taking solid assumptions. For instance the fact that every **Failure to achieve an Intended result p (def. 3.9)** implies a **Prevention of the achievement of the Intended result p (def. 3.4)** is provable only on the base of the following assumption.

4. The notion of Attemptß

After having presented the notions of Prevention and Failure we would like to arrive at the core notion of this analysis: the notion of attemptß. Again the conceptual nucleus of attemptß is the relation between the action and the outcome p to be achieved: action $a_1 \rightarrow$ outcome p .

Consider the following proposition:

“Agent x has attempted to kill agent y ” = “Agent x has attempted to do some action in order that agent y will be dead”.

Two different meanings can be assigned to the previous propositions.

1. agent x had the present-directed intention to “put poison in agent x ’s glass” relative to the future directed intention that “agent y will be dead”, agent x was not sure that “agent y will die” and finally agent x started to “put poison in agent x ’s glass”.
2. agent x had the present-directed intention to “put poison in agent x ’s glass” relative to the future directed intention that “agent y will be dead”, agent x was not sure that “if he puts now poison in agent x ’s glass” then “agent y will die” and finally agent x started to “put poison in agent x ’s glass”.

It is interesting to show that the statement “*if the agent had some doubt about the fact that the chosen action would have led to “agent y will die” then the agent had also some doubt about the fact that the intended outcome would have been achieved*” (he was not sure that “agent y would have died”)” is not valid whereas the inverse statement is valid. Indeed it is not always true that “having some doubt about the conditional $a_1 \rightarrow p$ implies having some doubt about the achievement of the outcome p ”. We can imagine the situation in which an agent has planned to execute a certain action “count sheeps ” in order to achieve the result “get asleep” having some doubt about the fact that he will effectively get asleep by counting sheeps. At the same time the agent could believe that there will be enough time to find whatever substitutive action such that p will be surely achieved by means of it. *In such previous case the agent is completely sure to achieve p even if he has some doubt about the fact that the action actually chosen for achieving p will bring about p .*

Several determinants of doubt in the mind of the agent (or the possibility of failure) can be identified.

Agent x may be unsure that a chosen action will bring about the intended result either because Agent x believes that the chosen action is not a perfect mean to achieve the final result or because he is not completely sure that the conditions for the execution of the action are holding. Also social schemata can affect the way agent x believes that a certain result is not achievable²⁰.

²⁰ All these determinants of the belief about possible failure are specific kinds of belief that play a role in the practical reasoning (beliefs about ability, beliefs about instrumentality, beliefs about necessary conditions and resources for executing an action). See Castelfranchi (1996) for the model of goal processing where eight phases of goal processing are identified and nine different categories of beliefs, goals (or intentions) play a role in each of them. Beliefs in each phase are *reasons* for selecting a certain motivation $G1$ rather than another motivation $G2$. In that model also beliefs about preferences and beliefs about conflicts between motivational states are identified. We could in principle conceive attempts with respect to those kinds of beliefs. We can assume that an agent has the meta-goal to maximize his own expected utility and the meta-goal to pursue those results that are the best ones. Whenever the agent execute an action in order to achieve a certain result and he is not sure that the pursued result is the most preferred one then the agent “is fully attempting to achieve the most preferred result”.

For summarizing, we can distinguish three important macro-categories of attempt.

Definition 4.1

Attempt to achieve an intended result p.

For each agent x and state of the world p if

1. It exists an action a1 such that agent x had a present directed intention “to do an action a1” relative to a future directed intention that “the state of the world p will hold after the execution of action a1”;
2. agent x had some doubt about the fact that “p will hold after the execution of action a1”.

then *Agent x has attempted to achieve p.*

Definition 4.2

Attempt to achieve an intended result p by doing the planned action a1.

For each agent x, state of the world p and action a1

1. agent x had a present directed intention “to do action a1” relative to a future directed intention that “the state of the world p will hold after the execution of action a1”;
2. Agent x had some doubt about the fact that “action a1 will bring about p” (i.e. Agent x has some doubt about the fact that “action a1 will happen next and p will hold at the end of it”);

then *Agent x has attempted to achieve p by doing the planned action a1.*

Definition 4.3

Attempt to execute the planned action a1.

For each agent x and action a1

1. It exists a state of the world p such that agent x had a present directed intention “to do action a1” relative to a future directed intention that “the state of the world p will hold after the execution of action a1”;
2. Agent x had some doubt about the fact that “action a1 will be executed next”;

then *Agent x has attempted to execute action a1.*

Till now we have characterized the following statements:

Agent x has attempted to achieve p;

Agent x has attempted to achieve p by doing the planned action a1;

Agent x has attempted to execute action a1.

Given the previous definition Attempts can be associated either with success or with failure that is all the following cases are conceivable

1. *Agent x has attempted to achieve p and he has failed to achieve it.*
2. *Agent x has attempted to achieve p and he has succeeded to achieve it.*
3. *Agent x has attempted to achieve p by doing the planned action a1 and he has failed to achieve p by doing a1.*

4. *Agent x has attempted to achieve p by doing the planned action a1 and he has succeeded to achieve p by doing a1.*

5. *Agent x has attempted to execute action a1 and he has failed to execute it.*

6. *Agent x has attempted to execute action a1 he has succeeded to execute it.*

The three following kinds of Attempt are also quite general notions that should be characterized.

7. *Agent x is attempting to achieve p;*

8. *Agent x is attempting to achieve p by doing the planned action a1;*

9. *Agent x is attempting to execute action a1.*

The previous cases 7-8 of Attempt share with previous cases 1-6 the structure of mental states but cases 7-9 unlike cases 1-6 are judgments about actual attempts of the agent under observation. Those judgments are expressed while the agent under observation is actually executing the action that he has planned for achieving the intended result having some doubt about the fact either that the intended result p will hold α that the chosen action will bring it.

Given the analysis of Prevention and Failure that we have provided, we have all the tools to model Attempt in a more complete way. Indeed

by assuming that

every time an agent has a present-directed intention or a future-directed intention he knows to have it

it follows that

every attempt implies a Belief about a future Failure and more precisely

If agent x has Attempted to achieve the intended result p (def. 4.1) then agent x had a Belief about a future Failure to achieve the intended result p (def. 3.9)

If agent x has attempted to execute the planned action a (def. 4.3) then agent x had a Belief about a future Failure to execute the intended action a (def. 3.10)

If agent x has attempted to achieve an intended result p by doing the planned action a (def. 4.2) then agent x had a Belief about a future Failure in the planned achievement of the Intended result p (either def. 3.11.a or 3.11b or 3.11c).

As shown in paragraph 3.2 under particular assumptions the logical equivalence between the notion of Failure and the notion of Prevention can be proved. It follows that under those assumptions also the logical consequence that goes from the Attempt to the Expectation about a Possible Prevention can be established.

4.1 What do expecting (prospecting) the failure and having a doubt about the success mean?

Since on the basis of the Cartesian Argument it is impossible being sure that a certain event will happen in the future (we can be sure only about facts we directly perceives in the present, i.e. only the beliefs determined by perceptive sources can have maximum strength) it seems that all decision of executing a certain action and all executions of actions are attempts. This type of mistaking argument has been previously noticed by O'Shaugnessy (1972).

From an objective point of view all actions could fail so whatever action is an attempt STATEMENT 1

The statement is a direct consequence of the following statement.

It is always possible to find reasons for believing that a certain action will fail STATEMENT 2

We think that O'Shaugnessy's argument is correct only for attempts in an objective sense. Indeed, there could be cases in which agent does not prospect failure when doing a certain action. It is much more correct to say

From a subjective point of view an agent is attempting if and only if he prospects a possible failure

It follows that not necessarily every decision concerning a given action is equivalent to an attempt.

But what does really mean that an agent is not prospecting failure? We think that two different situations should be identified.

1. The agent has a certain implicit belief about failure: the belief is potentially derivable from the base of knowledge of the agent but it is not actually derived (for a definition of Implicit Belief in this sense see Cohen & Levesque, 1984 and Fagin & Halpern 1987) that is actually similar to the idea that the agent is not cognitively framed on the possibility of failure (Tversky & Kahneman, 1981). Besides, the notion of potential derivability depends on the identification of an external observer who has access to the knowledge base of the agent.
2. The agent has an explicit belief about failure but that belief has null strength. On the other side the belief about Success has a degree of strength x that is different from value zero. We distinguish here two sub-cases:
2a. the agent believes that all evidences that *should* be taken into account to make a judgments have been questioned. Moreover, the agent is completely sure that he will be successful (the belief about Success has a maximal strength); indeed the whole complete and available information that he has considered to make the judgment supports the belief about Success and he does not have on the other side any reason to believe that he will fail. From a point of view of an external observer who takes the previous statement 2 as a valid one

the agent is in a state of *objective ignorance*. Indeed according to the observer there are many more sources that *should* be taken into account to make a judgments and that are not actually considered by the agent. According to the observer (who agrees with statement 2) those additional sources would bring arguments in favor of failure.

2b. The agent believes that some evidence that *should* be taken into account to make a judgments has not been questioned. But given the incomplete available information the agent has some reason to believe that he will be successful and he does not have on the other side any reason to believe that he will fail. From his own point of view, the agent evaluates himself to be *ignorant* (let us call it **subjective ignorance**). According to the agent there are many more sources that *should* be taken into account to make a judgments and that he has not actually considered.

After having specified the different situations in which an agent does not prospect a possible failure it basically relevant for our analysis to specify what it really means that the agent has some doubt about the fact that he will succeed. To understand that is in fact necessary in order to have a full clarification of the notion of Attemptß.

The notion of ignorance is again a crucial notion for distinguishing the three situations in which the agent has some doubt about the fact that he will succeed. Indeed according to our view both Uncertainty and Ignorance play a role in the determination of the degree of doubt (Perplexity). We argue here that the three following situations are representative of situations in which the agent has some doubt about the success.

1. the agent believes that all evidences that *should* be taken into account to make a judgments have been questioned. Moreover, the agent is not sure that he will be successful; indeed some available information that he has considered to make the judgment supports the belief about Failure (UNCERTAINTY ABOUT SUCCESS + NO IGNORANCE).

2. The agent believes that some evidence that *should* be taken into account to make a judgments has not been questioned. But given the incomplete available information the agent has some reason to believe that he will be successful and some reason to believe that he will fail. Again from his own point of view, the agent evaluates himself to be *ignorant* (again **subjective ignorance**) but even if he evaluates himself to be ignorant the agent has some reason to believe that he will fail Failure (UNCERTAINTY ABOUT SUCCESS + IGNORANCE).

The following situation that is identical to previous situation 2b is quite particular since the agent does not prospect the failure but he has a doubt (exclusively determined by the ignorance) about the fact that he will succeed.

3. The agent believes that some evidence that *should* be taken into account to make a judgments has not been questioned. But given the incomplete available information the agent has some reason to believe that he will be successful and he does not have on the other side any reason to believe that he will fail. From his own point of view, the agent evaluates himself to be *ignorant* (let us call it **subjective ignorance**) but even if he evaluates himself to be ignorant the agent has some specific reason to believe that he will fail (CERTAINTY ABOUT SUCCESS + IGNORANCE).

In the next paragraphs we want to present briefly our general theory of Ignorance. We will present afterwards the notion of Attempt ? and two general notions of Trying and Hazarding that are for us sub-categories of the Attempt category.

5.2 Ignorance: a cognitive approach

Strength of belief, uncertainty, and ignorance (or ambiguity) are all epistemic dimension. We want to focus in this paragraph on the second dimension.

The concept of ignorance has been extensively investigated both in psychological and economics (Shackle, 1972, Ellsberg 1961, Smithson 1994) and formal approaches to quantify it have been proposed (see Halpern 2003 for a review). Moreover ambiguity has been shown to affect the decision process and ambiguity aversion in subject has been identified and formalized in different ways (see Camerer & Weber, 1992 for a review). We briefly present here our cognitive approach concerning Ignorance and Ambiguity (see Pezzulo et al. 2004 for the general model).

We claim that Ignorance is a subjective evaluation of actual lack of information on the basis of cognitive *evidential models* (the agent believes that certain evidences *should* be gathered in specific domains of decision) *reference class dependent*: an agent has a model (script) of his sources that allows him to evaluate that a certain type and a certain number of sources can provide *sufficient information* for reducing ignorance close to zero with respect to a given reference class²¹. In this way the strength of the belief and the (perceived) ignorance are two different measures, the second belonging to the meta-level.

Intuitively ignorance depends on how much information I have with respect to how much it exists; in an open world there is a potentially infinite number of witnesses that have not been questioned; so if we calculate ignorance in this way the agent has always the maximum degree of ignorance. *The agent does not know how many witnesses he can consider at most or better he does not know how he can reduce his ignorance close to zero.* A qualitative and cognitive analysis is here required. Here we shift the issue to an evidential and subjective level²². At this point the notion of *Structure of Classes of Acceptable Ignorance* (SCAI) is introduced.

²¹ With the term reference class (Gigerenzer et al., 1991) such as weather forecasting or football matches result forecasting, we mean here a domain of judgment.

²² Our notion of ignorance is very close to the notion of ambiguity identified in some recent economical and psychological literature where is stressed that decision making is affected by the decision maker's evaluation of his or her actual available information and competence to make judgments in specific domains (Heat & Tversky 1991). Instead, our approach is quite far from the notion of Sample Space Ignorance given in Support Theory (Tversky & Koehler 1994) where it is claimed that people do not follow the extensional logic of conventional probability theory. In Support Theory an agent can actually "ignore" actual information in the sense that he is not explicitly evaluating that evidences concerning a certain event e1 are also evidences concerning another event e2. Indeed it has been shown that unpacking (making information available for explicit evaluation) a compound event into disjoint components tends to increase the perceived likelihood of that event. An immediate implication is that unpacking an hypothesis and/or repacking its complement will increase the judged likelihood of that hypothesis.

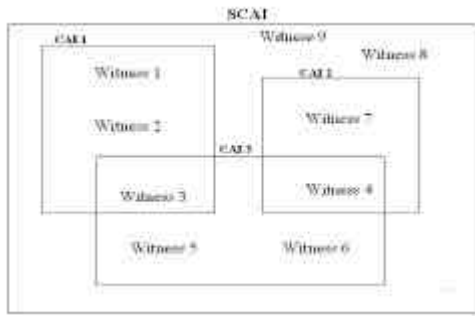


Figure 5

Each agent has a SCAI that includes several *Classes of Acceptable Ignorance* (CAI) that include one or more sources (e.g. Witnesses), each having its reliability value. For instance, $CAI_1 = (\text{witness 1, witness 2, witness 3})$ could be one of those classes. Classes of acceptable ignorance can be intersected and unified (see Fig. 1): they have the normal properties of sets in set theory.

The agent knows that testing all witnesses in a given class is enough for making the ignorance acceptably close to value zero. Imagine for example that the agent wants to know if tomorrow will rain or will be sunny. He has several classes of acceptable ignorance. For instance he can believe that by acquiring information about tomorrow's weather from source 1 = "New York Times" and source 2 = "CNN" is enough for making ignorance acceptable. Moreover, the following points are crucial for understanding how the relation between SCAI and agents works.

The agent has explicit models (meta-level) of Classes of Acceptable Ignorance as shown in Fig.1. There are witnesses who are included in classes of acceptance but also witnesses who are not included in any class.

Class-Ignorance is given for each class at a certain point of a query sequence (q_1, \dots, q_n) and is defined as the total number of witnesses in the class *minus* the number of tested witnesses in that class, weighted for the inverse of the total number of witnesses in the class.

Absolute-Ignorance is defined as the minimal value of Class-Ignorance among all CAIs.

Class-Ignorance (Class n, q_i) =

$$(n.\text{wit. (Class n, } q_i) - n.\text{queried.wit. (Class n, } q_i)_{\text{Agent } x, q_i}) / n.\text{wit. (Class n, } q_i)_{\text{Agent } x, q_i}$$

Absolute-Ignorance (q) =

$$\text{Min}_{\text{Class } x} (\text{Class-Ignorance (Class } x, q_i)_{\text{Agent } x, q_i})$$

Let us finally remark the difference between Ignorance and **Uncertainty**.

Uncertainty in our view is a measure of the difference between the value of strength of the belief that p (for example the belief about success) and the value of strength of the belief that not p (for example the belief about failure). When the difference is 0 the value of uncertainty is maximal, when the difference is 1 the value uncertainty is minimum.

5. Attempt ? : on the side of the observer

In the previous analysis of Attempt β the belief about the possible failure in the achievement of the final intended result was in the mind of the agent who executed the action. Let us assume here that the doubt is in the mind of an external observer. Attempts ? can be defined as follows.

Definition 5.1

Observed attempt to achieve an intended result p.

For each agent x and state of the world p if

1. It exists an action a1 such that agent y believes that agent x had a present directed intention about “to do action a1” relative to a future directed intention that “the state of the world p will hold after the execution of action a1”;
2. Agent y had some doubt about the fact that “p will hold after the execution of action a1”.

then *According to agent y agent x has attempted to achieve p.*

Definition 5.2

Observed attempt to achieve an intended result p by doing the planned action a1.

For each agent x, state of the world p and action a1

1. agent y believes that agent x had a present directed intention “to do action a1” relative to the future directed intention that “the state of the world p will hold after the execution of action a1”;
2. Agent y had some doubt about the fact that “action a1 will bring about state of the world p” (i.e. Agent y had some doubt about the fact that “action a1 will happen next and p will hold at the end of it”);

then *According to agent y agent x has attempted to achieve p by doing the planned action a1.*

Definition 5.3

Observed attempt to execute the planned action a1.

For each agent x and action a1

1. It exists a state of the world p such that the agent y believes that agent x had a present directed intention “to do action a1” relative to a future directed intention that “the state of the world p will hold after the execution of action a1”;
2. Agent y had some doubt about the fact that “action a1 will be executed next”;

then *According to agent y agent x has attempted to execute action a1.*

Finally there are the extreme cases in which the external observer agent y attributes the belief about failure to agent x who is the agent of the attempt (those cases represent full attributions of attempt).

6. Attempting and Trying

We want to make in this paragraph a preliminary distinction between what till now we have referred to, that is the verb *to attempt* and a similar common sense verb: the verb *to try*.

We think that the distinction between Attempting and Trying has not been very clarified in the recent philosophical literature about Individual Action theory.

Many philosophers have debated about the following issue:

Is it possible to try to do something without intending to do it?

For instance Bratman (1987) have given a positive answer to the previous question. After having illustrated the target games Bratman says.

My response is to reject the contention that the agent must intend to each target. What the agent needs to do is to try to hit each target. But this does not mean that the agent must intend to hit each target...If the agent nevertheless does intend to hit each target, the agent is criticizably irrational (pag. 117)

The presentation of the target game is related with a criticism to the Simple View. According to Bratman the agent in the target game cannot intend to hit both target given the requirement of rational consistency for the intentions. The agent in the target game merely intends to try to hit both targets without intending to hit both targets and a “guiding desire or goal” to hit both targets plays the role that is normally played by the intention in guiding the action²³. The position of Bratman with respect to the target game have been criticized by several authors. For instance Tuomela (1995) proposes the idea that in the scenario the agent has a disjunctive intention (the intention to hit target 1 or target 2) and that intention has a role in guiding the action; Mele (2003) on the other side assigns to the “intention to try to hit both targets” the role in guiding the action and refuses the idea concerning the “guiding desire”. Finally McCann (1991) and Adams (1997) completely refuses the idea that an agent *can try to do something without intending to do it*. McCann defends the Simple View by arguing that: 1) it does not make sense to introduce a further notion such as the notion of guiding desire with the same properties of the intention mental state; 2) trying to do A cannot be considered as an action. We agree with the second McCann’s objection: [trying to do an action a] cannot be considered as an action and so it can not be the content of a certain intention. Moreover, in the next paragraph we will provide an alternative way to deal with the issue raised by Bratman by giving a definition of “Trying to do some action a1” that does not imply the intention to do the action a1.

²³ In his criticism towards the Simple View Bratman proposes the “Single Phenomenon View” according to which “to A intentionally I must intend to do something that it is not necessarily A”.

6.1 The notion of Trying

Our main thesis are the following ones.

1) When an *agent is trying to do some action a1* the object of the present (primary) intention is merely epistemic that is the primary intention is the *intention to know whether action a1 will be successful in the achievement of a certain result.*

2) When an *agent is trying to do some action a1*, the agent is actually intending to do either action a1 or some sub-component of action a1.

Starting with the previous two thesis we can distinguish three different kinds of Trying.

Definition 6.1

Trying to achieve an intended result p.

For each agent x and state of the world p if

1. It exists an action a1 such that agent x had a present directed intention to do [a part of] action a1 and simultaneously to check whether p will hold at the end of [the part of] a1 in order to know whether p will hold at end of a1 [in order to know whether if a1 is executed next, p will hold at the end of it];
2. agent x had some doubt about the fact that “if a1 is executed next, p will hold at the end of it”;

then *Agent x has tried to achieve p.*

Definition 6.2

Trying to achieve an intended result p by doing the planned action a1.

For each agent x, state of the world p and action a1

1. agent x had a present directed intention to do [a part of] an action a1 and simultaneously to check whether [the part of] a1 happens next and to check whether p will hold at the end of it in order to know whether a1 happens next and p will hold at the end of a1 [to know whether if a1 is executed next, a1 happens next and p will hold at the end of it];
2. agent had some doubt about the fact that if a1 is executed next, a1 happens next and p will hold at the end of it;

then *Agent x has tried to achieve p by doing the planned action a.*

Definition 6.3

Trying to execute the planned action a.

For each agent x and action a1

1. It exists a state of the world p such that agent x had a present directed intention to do[a part of] an action a1 and simultaneously to check whether [the part of] a1 happens next in order to know whether if a1 is executed next, a1 happens next;
2. Agent x had some doubt about the fact that if a1 is executed next, a1 happens next;

then *Agent x has tried to execute action a1*.

In defining Trying with respect to the execution of *a part of* the full action we have implicitly assumed the existence of particular inferential processes based on the existence of the following epistemic rules.

- **Trying to achieve an intended result p:**

Agent x believes that if p holds at the end of the part of action a1 then if a1 was executed next, p would have held at the end of it.

- **Trying to achieve an intended result p by doing the planned action a1.**

Agent x believes that if p holds at the end of the part of action a1 then if a1 was executed next, a1 would have happened next and p would have held at the end of it.

- **Trying to execute the planned action a1.**

Agent x believes that if the part of action a1 is executed next then if a1 was executed next, a1 would have happened next.

From the previous definition it follows that “Trying to do a1” does not necessarily implies a present directed intention to do a1. It simply requires the intention to do *a part* of it. In the target example discussed by Bratman in order that the agent tries to hit both targets (action a1), it is simply required that the agent has a present directed intention to do **a part of** action a1 relative to the future directed intention to know whether the action a1 will be successful. Given our model of Basic and Complex Actions (see paragraph 2.2) a part of an action can be identified inside a causal complex of movements of the body and delegated external events. Imagine the agent has action a1 as an object of a present-direct intention where a1 can be decomposed in distinct sub-components: “raise both hands”, “direct the left hand on the left button”, “direct the right hand on the right button”, “press the left button with the left hand” and simultaneously “press the right button with the right hand” that are all movements of the body that the agent intends to execute; “the electric signal goes from the left button to the left target”, “the electric signal goes from the right button to the right target”, “the left target shoots down”, “the right target shoots down” that are all delegated external events. Even if the agent believes that action a1 cannot be realized and so he cannot (by the rationality assumption) intend to do it, he can try to execute that action by planning to execute a part of a1 and to check whether the part of a1 happens next and finally in order to know whether if a1 was executed next, a1 would have happened next. The part of action a1 that the agent decides to execute and the related checking action could be for instance the “left side of action a1” and the simultaneous checking whether the execution of the “left side of action a1”.

The execution of the “left” sub-sequence of action a1 and the simultaneous check whether about the successful execution of “left side of action a1” could be considered by the agent as a mean for having an additional confirmation of the fact that “if a1 was executed next, a1 would not have happened next”.

Imagine finally an agent who is trying to open the door. He can simply stop the sequence of bodily movements before the door is open and still trying to open it. The execution of the half sequence of bodily movements and the contemporary test on the execution of the action could be enough according to agent as a

mean for knowing whether if the complete action of opening the door was executed next, “opening the door” would have happened next. There are many real life situations in which we merely execute a part of an action in order to know whether if the complete action was executed next, the complete action would have happened next.

Notice that the kind of argument we propose here for justifying the statement

It is possible to try to do something without intending to do it

does not apply to attempt i.e. the next statement is not valid in our view

It is possible to attempt to do something without intending to do it.

Indeed as shown in the previous definitions of Attempt

necessarily if an agent attempts to some action a1 then the agent intends to do that action.

In the next paragraph we will describe more carefully the epistemic component of every Trying. In order to provide this description we will briefly review the notion of Epistemic Action.

6.2 The real core of Trying and the problem of the Epistemic Action

“Epistemic Actions” (EpAs) are *actions aimed at acquiring knowledge from the world*; any act of active perception, monitoring, checking, testing, ascertaining, verifying, experimenting, exploring, enquiring, give a look to, etc. In our previous work (Lorini & Castelfranchi 2004) we have provided a general taxonomy of EpAs. First all we have shown that Epistemic Action are done in order to achieve an Epistemic Goal (or an Epistemic Intention) where an Epistemic Goal can be represented as the goal to “know the truth value of certain proposition”.

We have classified EpAs according to two dimensions.

- Modal dimension: how a specific epistemic action is realised.
- Functional dimension (why are EpAs performed? Which is their purpose?): the kinds of belief that the EpA is aimed at verifying (beliefs about ability, beliefs about instrumentality, beliefs about opportunity etc...each playing a specific role in the goal processing).

We report here only some results concerning the modal dimension. We have distinguished two general macro-modal categories of epistemic actions:

- *Pure Epistemic Actions* (Waiting for);
- *Parasite Epistemic Actions* (Looking for).

Pure epistemic actions are identifiable through verbs of sensory actions: to see, to observe, to hear etc...We define them as: *actions that are specialised for epistemic functions: to the acquisition of knowledge, the verification (confirmation) of beliefs etc...* Since actions are identifiable in terms of the results they are done for, we could define pure epistemic actions as: *Actions that are specialised for achieving epistemic results: to the acquisition of knowledge, the verification (confirmation) of beliefs etc...* Pure epistemic actions never

change the state of the world, they just change the knowledge of the agent (see Herzig et al. 2000 Van Linder et al. 1994).

Parasite epistemic actions on the other side are: *pure epistemic actions that exploit pragmatic actions in order to achieve the epistemic result they are done for*. Parasite epistemic actions change both the state of the world and the knowledge of the agent. We have identified two forms of parasitism of pure epistemic actions on pragmatic ones. Pragmatic actions and epistemic actions can be done *sequentially* or in *parallel*. The pragmatic action is a mean for achieving the epistemic goal (to know the truth value of p). In order to know whether p is true or not the agent has to check whether p is true or not (pure epistemic action) and before it the agent has to execute a pragmatic action $a1$. The pragmatic action (the first component of the plan) creates the conditions for the execution of the pure epistemic action (second component of the plan). For example, imagine an agent wants to know whether “outside it is snowing or not” (epistemic goal). The agent will first open the door of the house (action $a1$) and will move his own attention towards the external environment (pure epistemic action).

Moreover we have addressed another general distinction:

- *Proposition-based epistemic actions* (Check whether). I test the truth value of a specific proposition;
- *Proposition-free epistemic actions* (Check what).

Finally we distinguished EpAs depending on the sources of confirmation (of test):

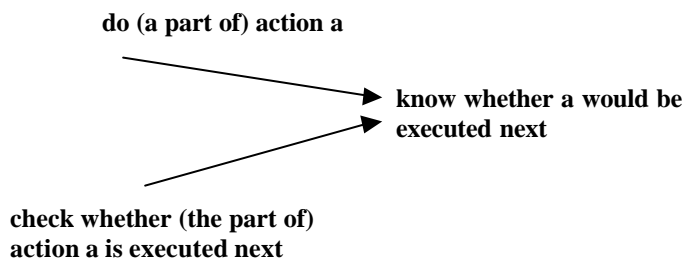
- Pure Epistemic Actions on perceptive sources;
- Pure Epistemic Actions on inferential sources.

An agent executes a Pure EpA on a perceptive source in order to test a belief that p if and only if he is simply testing the belief that p “by matching the epistemological representation of the object designated in p with the perceptual object (the referent) and no inferential process is involved in the process of verification”²⁴. On the other side an agent executes an EpA (either Pure or Parasite) on an inferential sources in order to test a belief that p if and only if he is testing the belief that p through some inferential process and no match with perceptual object is involved in the process of verification²⁵.

Given a model of Epistemic Action the cognitive structure of a Trying becomes clearer. Take for instance the definition of **Trying to execute the planned action a** (definition 6.3). The following figure shows the structures of the mental plan (a Parasite Epistemic Action) for achieving the Epistemic Goal of “knowing whether action a would be executed next”.

²⁴ This way to approach to the problem is similar to the one given in (Minsky 1974) where the process of testing through perception the appropriateness of a Frame (for representing categories and situations) is analyzed. We have assumed that a proposition p can be specified as a predicate of the following form “the |perceived object| is a |dog|” where the category |dog| can be organized (as in Davidsson, 1993) in terms of an *epistemological representation* (which is used to recognize through perception the instances of a category, for instance a 3D model) and an *inferential representation* (collection of encyclopedic knowledge about the category) and a *designator* (the symbol used to refer to the category).

²⁵ We did not consider in the present analysis *Pure Epistemic Actions on mnestic sources* that are according to Audi (2001) another important kind of Basic sources of belief (together with perceptive sources) where Basic Sources of



Our notion of Trying has all the features identified by Von Wright (1963). Indeed we agree with the following ideas:

...It would be a mistake to think that whenever an agent has successfully accomplished an act he has also tried to accomplish it [...] To construe every act as a result or consequence of trying to act would be a distortion... (pag. 51).

Our definition of Trying also agrees with the following statement

...In the course of trying to do something, one may perform various acts. Basically trying seems to me to belong to the category of activity. Trying to do something may, as we say, 'result' in the act's being successfully performed. But performing the act is not tied to trying to perform it in the same way as the resulting change is tied to the doing of the act...(pag. 52).

Indeed every Trying (as we defined it) implies (the intention to do) a Parasite Epistemic Action and so every Trying implies (the intention to do) a Complex Action (see paragraph 2.2) composed at least by two active contributions of the agent: The execution of the action + The test on the execution of the action. Since the Epistemic Test could be decomposed in several pragmatic sub-components (for instance “turn the head”, “fix the attention” etc...) it follows (in accordance with Von Wright) that in many cases a Trying implies the execution of various pragmatic acts (in our definition those acts are aimed at achieving an Epistemic Goal).

After having presented the notion of Epistemic that is a crucial notion for the conceptual analysis of the Trying we want to answer in the next paragraph to the following question “Why should an agent try to do a certain action?” by means of the analysis of Discovery Learning.

6.2 The case of discovery learning: Projecting the Trying towards the future

Interesting schemas of *discovery learning* (see Simon & Anzai 1979 for further analysis) can be designed starting from the present analysis of Trying. Discovery learning is in fact the intentional version of functional learning that was already instantiated in TOTE unit (see Miller et al. 1961) and that is embedded in BDI systems. In BDI models and in TOTE at each round after the execution of an action the agent tests

belief are those experiential sources whose justificatory power is not derivative (it is not obtained by some inferential

automatically whether the action has been successful in the achievement of a given result. If the action a_1 has been successful then the belief about the fact that p is true and the belief about instrumentality $a_1 \rightarrow p$ is strengthened (viceversa in case of failure) as well as beliefs about ability etc... That test is not intentional as well as the learning associated with it: the test is automatic and the learning is merely functional. On the contrary in *discovery learning* the agent has the explicit epistemic goal G_3 of discovering (learning) a good procedure to achieve a pragmatic intended result p at time t_1 and instrumental to G_3 the agent has the explicit goal of

testing whether a given procedure a_1 is a good mean for achieving result p at time t_1 .

The verification of the general belief about the “degree of instrumentality” (or “degree of correctness”) of the mean is inserted in a general plan of intentional learning for achieving the final intended result p . Indeed the agent could reason as follows. In order to achieve p at time t_1 we must discover first of all whether a procedure a_1 is effectively good for achieving p at t_1 . Let us assume that *the agent believes that if the procedure a_1 allows to achieve p at t_1-n then the procedure a_1 allows to achieve p at t_1 .* Given that it is much more economical for the agent to perform procedure a_1 at t_1-n and check whether a_1 is a good mean achieving p at t_1-n in order to know whether a_1 is a good mean achieving p at t_1 . In order to test his hypothesis about the future (at time t_1) correctness of action a_1 with respect to the result p the agent “try” to **achieve an intended result p at time t_1-n by doing the planned action a** (see above the **Definition 6.2 of Trying**).

In Discovery Learning situations the plan of “Trying” aimed at knowing whether a_1 is a good mean achieving p at t_1 is most of the time inserted in a more general plan of Discovery as shown in the following figure.

PLAN OF DISCOVERY

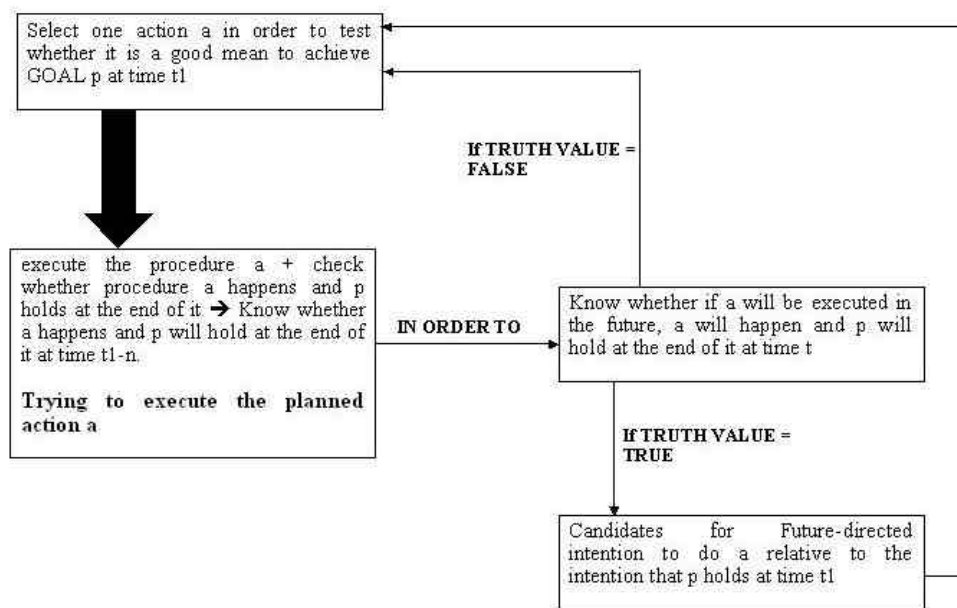


Figure 6

7. Subjective definitions for Hazardizing

We want to give here a definition for Hazardizing. Indeed that notion has some relation with the notion of Attempting. We want here to show differences and similarities.

Before analyzing Hazardizing we need a general definition for risk and threshold of acceptance for risk.

We call *expected risk as regard an action a1*²⁶ the *sum expected costs + expected failure* associated with an action a1

(the product of the value of final intended result with the strength of the belief about the fact that the final result will not be achieved).

We call the *threshold of risk acceptance*

The maximum level of risk as regard an action that the agent can accept before deciding to choose that action.

A procedural rationality (see Simon, 1990) version of Hazardizing for agent x as regard action a1 is the following.

²⁶ The common sense of risk is the following “*possibility of meeting danger or suffering harm, loss etc.*” (Longman Dictionary). In micro-economy the basic definition of risk is given by Knight (1921) where “*risk*” refers to situations where the decision-maker can assign mathematical probabilities to the randomness which he is faced with whereas “*uncertainty*” refers to situations when this randomness “cannot” be expressed in terms of specific mathematical

1. Agent x has an intention to do action a1 among a set of n actions relative to the Intention that p holds;
2. Agent x evaluates that action a1 overcomes the threshold of risk acceptance.

We should also include the following condition: agent x evaluates that action a1 leads to a relatively high value of estimated gains so to a quite high value of expected utility²⁷. Indeed it seems quite intuitive that hazarding implies a choice driven by the attraction for estimated rewards and a relatively high value of expected utility.

It is important to notice that the previous definitions of hazarding imply a deep reflection about the problem of rationality. In order to give a more efficacious definition for hazarding it is necessary to include ignorance in our argumentation.

In many situations agents do not have all information for assigning precise probabilities to the different options. *To hazard* could be realised as the choice of an action that the agent thinks to be risky given the information for believing and the consequent strengths of belief. The fact is that those values of risk (and also the value of expected utility) can be “engrossed” in ignorance, they are not completely precise.

We define ***interval of expected risks***

expected risk ± Ignorance about expected risk

The agent can be attracted by the idea of getting a lot of good rewards (perhaps a high value of expected utility) and decreases the strength of expected risk by “filling” the interval of ignorance in the appropriate way. When the fulfilment of the gap of ignorance is not based on epistemic reasons (new acquired information) the agent is hazarding.

It is also useful on this side to assume the existence of a threshold of ignorance acceptance in risk evaluation such that if the value of uncertainty is above the threshold then the agent can not accept the risk even if the assessed risk (based on the possessed knowledge) is very low. The ***threshold of ignorance acceptance in risk evaluation*** is defined as follows

We call the ***threshold of ignorance acceptance in risk evaluation***

*The maximum level of ignorance concerning risk related with action a1 that the agent can accept before proceeding to choose action a1*²⁸.

We can assume that the value of the threshold is fixed and does not vary depending on the value of risk i.e. we assume that the threshold is fixed as a norm (a rule) in the mind of the agent.

But the interesting case is when the value of the threshold can vary depending on the value of the risk.

Indeed even if the **expected risk** is below the threshold of risk acceptance the value of ignorance could not be acceptable i.e. the threshold of risk acceptance is internal to the interval of values of expected risks (higher is the ignorance wider is the interval). So we can say that the ***Ignorance about expected risk*** is above the **threshold of ignorance acceptance in risk evaluation** when

probabilities. Our definition is much closer to the common sense of Risk (we stress the negativity of the expected result).

²⁷ We can assume the existence of aspiration level of expected utility and the fact that the action overcomes it.

$Max(\text{interval of expected risks}) > \text{threshold of risk acceptance}$

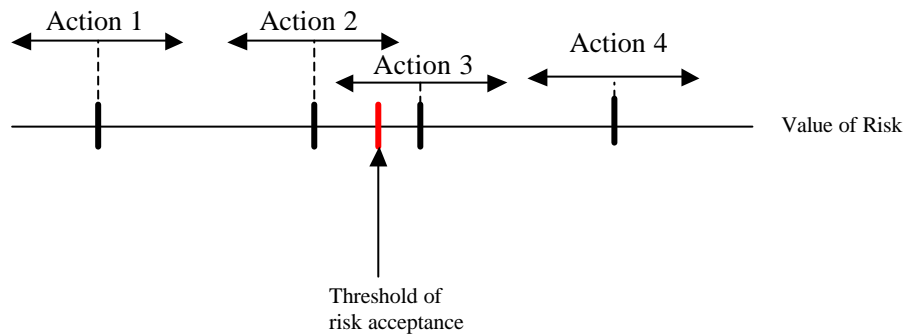


Figure 7

In figure 2 action 2, action 3 and action 4 have a level of ignorance about expected risk that is above the threshold of ignorance acceptance in risk evaluation. So if the agent chooses one of those actions is above the **threshold of ignorance acceptance in risk evaluation** so he is hazarding.

Let us suppose that the expected risk is below the threshold of risk acceptance and the Ignorance about expected risk is above the threshold of ignorance acceptance in risk evaluation (action 2 in fig. 2).

The agent can either try to reduce his ignorance or evaluating another action or executing the action. If he decides to execute the action he is weakly hazarding. If he decides of considering another action the process starts again.

If he decides to reduce ignorance several cases are possible. Let us assume that every update of knowledge, in the sense of taking new information, always reduces ignorance²⁹.

1. the expected risk is still below the threshold of risk acceptance and the ignorance about expected risk goes below the threshold of ignorance acceptance in risk evaluation;
2. the expected risk is still below the threshold of risk acceptance and the ignorance is still above the threshold of ignorance acceptance in risk evaluation;
3. the expected risk goes above the threshold of risk acceptance and the ignorance is still above the threshold of ignorance acceptance in risk evaluation.

After the update the process starts again. Let us suppose that the agent falls in previous condition 3. Again he can either try to reduce his ignorance or evaluate another action or execute the action. If he decides to execute the action is strongly hazarding.

In a dynamic of goal processing the two options of deciding to execute the action and deciding to consider another action are internal to the decision process. They can simply be described in a procedural (functional)

²⁸ The value of uncertainty in risk evaluation can be easily calculated as the mean value of the sum of uncertainty values concerning all considered costs and the failure in achieving the final intended result.

²⁹ In general it is plausible that taking new information “changes” the agent’s awareness about his own epistemic condition i.e. it can happen that after the updating the agent believes to be more ignorant he was before the update.

way. But the decision to reduce ignorance is always associated with the agent's evaluation of his own general epistemic state. A deep meta-level of reasoning plays a role. The agent is attracted by the fact that the action can lead to high value of expected rewards and expected utility. He "looks at" at his own actual knowledge and evaluates his own value of ignorance concerning the risk. So he decides (meta-decision) to execute a specific epistemic action in order to reduce that ignorance.

The introduction of a **threshold of ignorance acceptance in risk evaluation** allows to give several procedural and realistic definitions for hazarding.

Procedural definition of Strong Hazarding of agent x as regard action a1 (Action 3 in Fig.2)

Agent x is hazarding as regard action a1 iff

1. Agent x has an intention to do action a1 among a set of n actions relative to future-directed intention that p holds;
2. Agent x evaluates that action a1 overcomes the threshold of risk acceptance;
3. Agent x evaluates that Ignorance about expected risk related with action a1 is above the threshold of ignorance acceptance in risk evaluation.

Procedural definition of Weak Hazarding of agent x as regard action a1 (Action 2 in Fig.2)

Agent x is hazarding as regard action a1 iff

1. Agent x has an intention to do action a1 among a set of n actions relative to future-directed intention that p holds;
2. Agent x evaluates that action a1 does not overcome the threshold of risk acceptance
3. Agent x evaluates that Ignorance about expected risk related with action a1 is above the threshold of ignorance acceptance in risk evaluation.

When the agent evaluates that his own ignorance will never be "enough" for reducing the risk below the threshold of risk acceptance in case new information will be acquired (Action 4 in Fig.1), given formally $Min(interval\ of\ expected\ risks) > threshold\ of\ risk\ acceptance$ and decides to execute the action, he is really behaving irrationally.

Conclusion

We go back in this conclusion to show the relations between the category of Attempt α and Attempt β that we have discussed till now.

We will proceed by giving two conclusive statements and related arguments and explanations against them.

1) Every action a_1 starting at $t_1 \rightarrow$ previous present-intention to do a_1 at t_1

Apart from the discussion of paragraph 2.6 about Micro-Actions and Micro-Events, our analysis has neglected till now the subtle issue of automatic non-intentional actions. At the moment we accept that automatic non intentional merely goal-oriented³⁰ actions exist, the statement becomes invalid.

2) Every attempt α at $t_1 \rightarrow$ previous (future-directed) intention to do a_1 at t_1

The formula is invalid by assuming that automatic not intentional actions exist. This assumption contradicts the Psychophysical Law (see paragraph 1) given by O'Shaughnessy according to whom Trying is exclusively related with consciousness, is an intention-based notion.

The previous two statements and the related argumentation make explicit the fact that the notion of Trying should be investigate also at the automatic not deliberated layers of the behavior.

We leave this complicated issue to further analysis.

³⁰ For a distinction between goal-oriented and goal-directed behaviors see Conte & Castelfranchi (1995).

References

- Adams, F. (1997). Cognitive Trying. In G. Holmstrom-Hintikka & R. Tuomela (eds.) *Contemporary Action Theory*, Vol. I, 287-314, Dordrecht: Kluwer.
- Anzai, Y., & Simon, H. A. (1979). *The theory of learning by doing*. *Psychological Review*, 86, pp. 124-140.
- Audi, R. (2001). *The Architecture of Reason: the structure and substance of rationality*. Oxford University Press.
- Anscombe, G. E. M. (1957). *Intention*. Oxford University Press.
- Belnap, B. (1991). Backwards and Forwards in the Modal Logic of Agency. In *Philosophy and Phenomenological Research* vol. II, no. 4, 1991, pp 777-807.
- Belnap, B., Perloff, M. (1988). Seeing to it that: A Canonical Form for Agentives. *Theoria* 54, pp. 175--199.
- Brand, M. (2003). Activity and Passivity. In M. Sintonen, P., Ylikoski, K., Miller (Eds.), *Realism in Action*, Kluwer Academic Publisher.
- Bratman, M. E. (1987). *Intentions, plans, and practical reason*, Cambridge, MA: Harvard University Press.
- Camerer, Colin F., and M. Weber. (1992). Recent Developments in Modelling Preferences: Uncertainty and Ambiguity. *Journal of Risk and Uncertainty* 5, pp. 325-70.
- Castelfranchi, C. (1995). Representation and integration of multiple knowledge sources: issues and questions. In Cantoni, Di Gesu', Setti e Tegolo (Eds.), *Human & Machine Perception: Information Fusion*, Plenum Press.
- Castelfranchi, C. (1996). Reasons: Belief Support and Goal Dynamics. *Mathware & Soft Computing*, 3, pp. 233-47.
- Castelfranchi, C, Falcone, R. & Pezzulo, G. (2003). Trust in information sources as a source for trust: a fuzzy approach. *AAMAS 2003*: 89-96.
- Castelfranchi, C., Lorini, E. (2003). Cognitive Anatomy and Functions of Expectations. *IJCAI '03 Workshop on Cognitive modeling of agents and multi-agent interaction*, Acapulco, Mexico.
- Conte, R., Castelfranchi, C. (1995). *Cognitive Social Action*. UCL Press, London.
- Cohen, P. R. , Levesque, H. J. (1990). Intention is choice with commitment. *Artificial Intelligence*, 42, pp. 213-261.
- Davidson, D. (1980). *Essays on Actions and Events*. Oxford University Press.
- Davidsson, P. (1993). A framework for organization and representation of concept knowledge in autonomous agents. In *Scandinavian Conference of Artificial Intelligence*.
- Demolombe, R. (2004). Reasoning about trust: a formal logical framework. In *Proceedings of the 2d International Conference iTrust*. Oxford, 2004.
- Elgesem, D. (1997). The modal logic of agency. *The Nordic Journal of Philosophical Logic*, vol. 2, pp. 1-46.
- Ellsberg, D. (1961). Risk, Ambiguity and the Savage axioms. *Quarterly Journal of Economics*, 75, pp. 643-

669.

Fagin, R., Halpern, J. Y. (1988). Belief, Awareness, and Limited Reasoning. *Artificial Intelligence*, 34(1), pag. 39-75.

Falcone R., Castelfranchi C. (1998). Towards a Theory of Delegation for Agent-Based Systems. *Robotics and Autonomous Systems*, N° 24, pp.141-157.

Grosz, B.J., Kraus, S. (1996). Collaborative Plans for Complex Group Action. *Artificial Intelligence*, 86, pp. 269-357.

Fagin, R., Halpern, J. Y. (1987). Belief, awareness and limited reasoning. *Artificial Intelligence* 34, pp. 39-76.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98, pp. 506-528.

Halpern, J. Y. (2003). *Reasoning about Uncertainty*. The Mit Press.

Heath, C., Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty*, 4, pp. 5-28.

Herzig, A., Lang, J., Polacsek, T. (2000). A modal logic for epistemic tests. In *Proceedings of European Conference on Artificial Intelligence (ECAI'2000)*, Berlin, August.

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

Horty, J., Belnap, N. (1995). The deliberative stit: a study of action, omission, ability, and obligation. *Journal of Philosophical Logic* 24, pp. 583-644.

Kenny, A. (1975). *Will, Freedom and Power*. Basil Blackwell, Oxford.

Knight, F. H. (1921). *Risk, Uncertainty, and Profit*. Hart, Schaffner & Marx; Houghton Mifflin Company, Boston.

Levesque, H. J. (1984). A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence (AAAI-84)*, pages 198-202, Austin, TX.

Lewis, D. (1986). Causation. In *Philosophical papers*, vol. 2, pp. 159-163.

Lewis, D. (1973). *Counterfactuals*. Harvard University Press, Cambridge, Massachusetts.

McCann, H. (1991). Settled Objectives and Rational Constraints. *American Philosophical Quarterly* 28, pp. 25-36.

Mele, A. R. (2003). Intending and Trying: Tuomela vs. Bratman at the video arcade. In M. Sintonen, P., Ylikoski, K., Miller (Eds.), *Realism in Action*, Kluwer Academic Publisher.

Mele, A. R. (1994). Intentional Action. *Noûs* 28, pp. 39-68.

Mele, A. R. (2003). Agents' Abilities. *Noûs* 37, pp. 447-470.

Mele, A. R., and P. K. Moser (1994). Intentional action. *Nous*, 28, pp. 39-68.

Miceli M., Castelfranchi C. (2002). The mind and the future: The (negative) power of expectations. *Theory & Psychology* 12, pp.335-366.

- Miller, G., Galanter, E., Pribram, K. H. (1960). *Plans and the structure of the behavior*. Rinehart & Winston, New York.
- Minsky, M. (1975). A framework for representing knowledge. In P.H. Winston (Ed.), *The Psychology of Computer Vision*, pp. 211-277. McGraw-Hill.
- O'Shaughnessy, B. (1973). 'Trying (as the Mental 'Pineal Gland')'. *Journal of Philosophy* 70, pp. 365-86.
- Pezzulo, G., Lorini, E., Calvi, G. (2004). How do I know how much I don't know? A cognitive approach about Uncertainty and Ignorance. In *Proceedings of 26th Annual Meeting of the Cognitive Science Society (CogSci 2004)*, Chicago, USA, 5-7 August, 2004.
- Pollack, M. E. (1990). Plans as Complex Mental Attitudes. In P.R. Cohen, J. Morgan, and M. E. Pollack, *Intentions in Communication*, MIT Press.
- Pörn, I. (1977). *Action Theory and Social Science: Some formal models*. Synthese Library 120, D. Reidel, Dordrecht.
- Rao, A. S., Georgeff, M. P. (1991). Asymmetry thesis and side-effect problems in linear time and branching time intention logics. In C. Rich, W. Swartout, B. Nebel (Eds.), In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, Morgan Kaufmann Publishers, San Mateo, CA.
- Rao, A. S., Georgeff, M. P. (1992). An abstract architecture for rational agents. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCAI-91)*, Morgan Kaufmann Publishers, Sidney, Australia.
- Santos, F., Carmo, J., Jones, A. (1997). Action concepts for describing organised interaction. In R. A. Sprague (ed.), *Thirtieth Annual Hawaii International Conference on System Sciences*, pp. 373-382.
- Searle, J. (1983). *Intentionality*. Cambridge University Press.
- Sellars, W. (1966). Thought and Action. In Keith Lehrer (ed.), *Freedom and Determinism*, Random House; New York, NY, pp. 105-39.
- Shackle, G. L. S. (1972). *Epistemics and Economics*. Cambridge: Cambridge University Press.
- Simon, H. (1990). Invariants of human behavior. *Annual Review of Psychology*, 41, pp.1-19.
- Smithson, M.J. (1994). Uncertainty. In V.S., Ramachandran (ed.), *Encyclopedia of Human Behavior*, New York: Academic Press.
- Tuomela, R. (1995). *The Importance of Us*. Stanford University Press, Stanford.
- Tversky, A. and Kahneman, D. (1981). The framing of decisions and psychology of choice. *Science*, 211, pp. 453-458.
- Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101, 547-567.
- W. van der Hoek and M.J.W. Wooldridge (2003). Towards a Logic of Rational Agency. *Logic Journal of the IGPL* 11:2, pp. 135 –160.
- Van Linder, B., van der Hoek, W., Meyer, J.-J. Ch. (1998). Formalising abilities and opportunities. *Fundamenta Informaticae*, 34, pag. 53-101.

Van Linder, B., van der Hoek, W., Meyer, J.-J.Ch (1994). Tests as Epistemic Updates. In A.G. Cohn (Eds.) *Proceedings of the 11th European Conference on Artificial Intelligence (ECAI'94)*, pp. 331-335, Wiley, Chichester.

Von Wright, G. H. (1971). *Explanation and Understanding*. Routledge & Kegan Paul, London.

Von Wright, G. H. (1972). *On so-called practical inference*. *The Philosophical Review* 15, pp. 39-53.

Von Wright, G. H. (1963). *Norm and Action: A Logical Enquiry*. Routledge & Kegan Paul, London.