Intentional agents in defense

Emiliano Lorini, Cristiano Castelfranchi

Intentional Agents in Defense

Emiliano Lorini and Cristiano Castelfranchi

Institute of Cognitive Sciences and Technologies-CNR
Via San Martino della Battaglia 44, 00185, Roma, Italy

Abstract. Multi-agent systems (MAS) should not be conceived as only cooperative. As open systems situations of concurrence, competition and conflict often arise. Starting from this perspective it is relevant not only pro-social interaction modeling, but also a theory of trust and monitoring, giving special relevance to issues of security and defense: how can an agent prevent that dangerous actions of other agents and dangerous events will frustrate his goals? In this paper some relevant concepts for a general model of defense in intentional agents are analyzed and formally specified. Moreover an ontology of defensive goals and defensive strategies is studied.

1 Introduction

Security is a matter of defense and protection. More precisely it is a defensive concept. In fact security means to be safe, not be exposed to damages and harms, to be in a completely safe, reliable and trustworthy environment where there are solutions for protecting ourselves from possible attacks and dangerous events.

A safe agent is either an agent who does not need to pursue defensive strategies in order to achieve his goals and to accomplish his tasks or an agent who is capable of blocking and contrasting possible attacks (viz. an agent having the abilities and opportunities to perform defensive actions) and who can exploit other agents, structures, artifacts and institutions in order to prevent, discourage and block possible attacks.

Obviously the former is a very implausible condition, since environments are always uncertain and dynamic, and agents generally act in social contexts where goals and interests are often divergent and conflicts can easily emerge [1].

Thus, we must conclude that: *a principled approach to security requires a careful analysis of defense*.

Since we believe that a general theory of defense is still missing, in this paper we will try to fill such a gap by taking the first steps towards the development of a formal model of defense and an ontology of defensive goals and defensive strategies. We will use a multi-modal logic of time and action and we will explictly model informational attitudes (beliefs and expectations) and motivational attitudes (goals and intentions) of agents.

The application of modal logic to the analysis of issues concerning security is not new in literature. For instance in [2, 3, 4, 5] specific epistemic logics, collectively referred to as authentication logics, have been proposed to deal with authentification issues. Such modal logics have been developed as tools for verifying the correctedness of

security protocols, where one wants to ensure that agents obtain certain knowledge over time and that ignorance of potential intruders persists over the whole run of a protocol.

Our objective in this paper is different from the objective of authors working on logics of authentification. We are mainly interested in providing a conceptual analysis of defense for agents who act according to their beliefs and motivations. Indeed we think that, due the strict theoretical connection between security and defense, models and methodologies of security would strongly profit by this kind of investigation. We believe that the framework of multi-agent systems and its formal models of intentional agency are the most suitable to develop such a kind of analysis.
More precisely, we will try to clarify the following points at a high level of abstraction.

- Under which conditions should an agent defend himself from someone else, that is, what should an agent expect, want, believe, etc... before deciding to pursue a defensive a strategy?
- Which are the main types of defensive strategies and how do defensive strategies vary depending on the context and situation of attack?

In this work we suppose that defenders are intentional agents with specific kinds of defensive goals and expectations of attacks. This is somehow a quite restrictive assumption. Indeed elementary reactive agents too can defend themselves.[1] A more general theory of defense should consider intentional defensive behaviors as well as functional defensive behaviors. For instance a possible restricted meaning of agent *i*'s escape is the act of agent *i* driven by *i*'s goal of changing his spatial location in order to avoid the impact with an object, event, other agent, etc... The notion of escape can be extended to cover functional behaviors of elementary agents where the intention to escape is substituted either by the designed function or by the function acquired through evolution or reinforcement learning.

2 A Logic of Defense

In order to formalize some relevant concepts in our ontology of defense we exploit a very simple modal logic of time, action and mental states. We call this logic $\mathcal{DL}$ (*Defense Logic*). $\mathcal{DL}$ is based on a combination of a fragment of linear temporal logic [6], a fragment of dynamic logic [7] and Cohen and Levesque's logic of beliefs and intentions [8].

In $\mathcal{DL}$ there are two modal operators Bel and $Goal$ for mental states. The former modal operator is a standard operator for beliefs [9] and expresses what a given agent currently believes.

The modal operator $Goal$ refers to goals of an agent. We suppose that goals are consistent (viz. an agent cannot decide to pursue two goals which cannot be achieved at the same time).

In the basic version of $\mathcal{DL}$ we cannot reason about conflicting goals and goals which are incompatible with actual beliefs.

The primitives of the logic are the following:

[1] Nevertheless there are defensive strategies such as dissuasion which have a specific intentional connotation.

- a set of agents $AGT = \{i, j, ...\}$;
- a set of atomic actions $ACT = \{\alpha, \beta, ...\}$;
- a set of propositional atoms $\Pi = \{p, q, ...\}$.

The set of well formed formulas φ, ψ of the language $\mathcal{L}_{\mathcal{DL}}$ is defined by the following BNF:

$$\varphi := p|\top|\neg\varphi|\varphi \wedge \psi|\,[i:\alpha]\,\psi|\bigcirc\varphi|Bel_i\varphi|Goal_i\varphi$$

where *p* ranges over Π and α ranges over *ACT*.

$Bel_i\varphi$ is read "agent *i* believes that φ" whereas $Goal_i\varphi$ is read "agent *i* has goal that φ". $\bigcirc$ is a standard *next* modal operator of temporal logic ($\bigcirc\varphi$ is read "φ is going to hold at the next state") whilst $[i:\alpha]$ is a standard operator of dynamic logic and $[i:\alpha]\,\varphi$ is read "φ holds after every occurrence of agent *i*'s action α". Hence $[i:\alpha]\,\bot$ expresses "agent *i* does not do action α". We use the following abbreviation: $\langle i:\alpha\rangle\,\varphi =_{def} \neg\,[i:\alpha]\,\neg\varphi$. Hence $\langle i:\alpha\rangle\,\varphi$ has to be read "agent *i* does action α and φ holds after this action". Finally$\langle i:\alpha\rangle\,\top$ has to be read "agent *i* does action α".

2.1 Basic Semantics

A model of $\mathcal{DL}$ is defined by a tuple $M =< W, R_{\bigcirc}, R^{att}, B, G, V >$ where:

- W is a set of worlds.
- $R_{\bigcirc}$ is a mapping $R_{\bigcirc} : W \longrightarrow 2^W$ associating sets of possible worlds $R_{\bigcirc}(w)$ to each possible world *w*.
- R^{at} is a mapping $R^{at} : AGT \times ACT \longrightarrow (W \longrightarrow 2^W)$ associating sets of possible worlds $R^{at}_{i:\alpha}(w)$ to each possible world *w*.
- *B* is a mapping $B : AGT \longrightarrow (W \longrightarrow 2^W)$ associating sets of possible worlds $B_i(w)$ to each possible world *w*. For each possible world *w* there is an associated set of possible worlds $B_i(w) \subseteq W$: the worlds that are compatible with the agent's beliefs.
- *G* is a mapping $G : AGT \longrightarrow (W \longrightarrow 2^W)$ associating sets of possible worlds $G_i(w)$ to each possible world *w*. For each possible world *w* there is an associated set of possible worlds $G_i(w) \subseteq W$: the worlds that are compatible with agent *i*'s goals.
- *V* is a mapping $V : \Pi \longrightarrow 2^W$ associating sets of possible worlds to propositional atoms.

2.2 Truth Conditions

- $M, w \models p \Longleftrightarrow w \in V(p)$.
- $M, w \models \neg\varphi \Longleftrightarrow$ not $M, w \models \varphi$.
- $M, w \models \varphi \wedge \psi \Longleftrightarrow M, w \models \varphi$ and $M, w \models \psi$.
- $M, w \models \bigcirc\varphi \Longleftrightarrow \forall w'$ if $w' \in R_{\bigcirc}(w)$ then $M, w' \models \varphi$.
- $M, w \models [i:\alpha]\,\psi \Longleftrightarrow \forall w'$ if $w' \in R^{at}_{i:\alpha}(w)$ then $M, w' \models \varphi$.
- $M, w \models Bel_i\varphi \Longleftrightarrow \forall w'$ if $w' \in B_i(w)$ then $M, w' \models \varphi$.
- $M, w \models Goal_i\varphi \Longleftrightarrow \forall w'$ if $w' \in G_i(w)$ then $M, w' \models \varphi$.

3 Axiomatization

We take the following complete axiomatization of our simple modal logic of time, action and mental states.

Table 1. Axiomatization

0. All tautologies of propositional calculus
1. $\bigcirc(\varphi \rightarrow \psi) \rightarrow (\bigcirc\varphi \rightarrow \bigcirc\psi)$
2. $\bigcirc\neg\varphi \leftrightarrow \neg \bigcirc \varphi$
3. $[i:\alpha]\,(\varphi \rightarrow \psi) \rightarrow ([i:\alpha]\,\varphi \rightarrow [i:\alpha]\,\psi)$
4. $\bigcirc\varphi \rightarrow [i:\alpha]\,\varphi$
5. $Bel_i(\varphi \rightarrow \psi) \rightarrow (Bel_i\varphi \rightarrow Bel_i\psi)$
6. $\neg(Bel_i\varphi \wedge Bel_i\neg\varphi)$
7. $Bel_i\varphi \rightarrow Bel_iBel_i\varphi$
8. $\neg Bel_i\varphi \rightarrow Bel_i\neg Bel_i\varphi$
9. $Goal_i(\varphi \rightarrow \psi) \rightarrow (Goal_i\varphi \rightarrow Goal_i\psi)$
10. $\neg(Goal_i\varphi \wedge Goal_i\neg\varphi)$
11. $Goal_i\varphi \rightarrow Goal_iGoal_i\varphi$
12. $\neg Goal_i\varphi \rightarrow Goal_i\neg Goal_i\varphi$
13. $Goal_i\varphi \rightarrow Bel_iGoal_i\varphi$
14. $\neg Goal_i\varphi \rightarrow Bel_i\neg Goal_i\varphi$
15. $Bel_i\varphi \rightarrow Goal_i\varphi$
Rules of Inference
R1. $\frac{\vdash\varphi \;\vdash\varphi\rightarrow\psi}{\vdash\psi}$ (Modus Ponens)
R2. $\frac{\vdash\varphi}{\vdash\bigcirc\varphi}$ ($\bigcirc$-Necessitation)
R3. $\frac{\vdash\varphi}{\vdash[i:\alpha]\varphi}$ ($[i:\alpha]$-Necessitation)
R4. $\frac{\vdash\varphi}{\vdash Bel_i\varphi}$ (Bel_i-Necessitation)
R5. $\frac{\vdash\varphi}{\vdash Goal_i\varphi}$ ($Goal_i$-Necessitation)

Axiom 1 and rule of inference R2 define a minimal normal modal logic for the temporal operator $\bigcirc$. Axiom 2 expresses the interpretation of $\bigcirc$ by a total function:

- for every $w \in W$ if $w' \in R_{\bigcirc}(w)$ and $w'' \in R_{\bigcirc}(w)$ then $w' = w''$ and for every $w \in W$, $R_{\bigcirc}(w) \neq \emptyset$.

Axiom 3 and rule of inference R3 define a minimal normal modal logic for the operator $[i:\alpha]$.

Axiom 4 is a connection axiom time-attempt. A similar axiom concerning the connection between time and action has been studied in [10, 11]. The semantic counterpart of axiom 4 is:

- $R^{at}_{i:\alpha}(w) \subseteq R_{\bigcirc}(w)$.

Thus the set of worlds which are accessible from world *w via* an attempt to do action α is a subset of the set of next-worlds.

Axioms 5 and 9 with rules of inference R4 and R5 define a minimal normal modal logic for the operators Bel_i and $Goal_i$. Axioms 6, 7, 8, 10, 11, 12 express the interpretations of B_i and G_i by serial, transitive and euclidean functions:

- *Seriality* of B_i: for every $w \in W$ $B_i(w) \neq \emptyset$
- *Seriality* of G_i: for every $w \in W$ $G_i(w) \neq \emptyset$.
- *Transitivity* of B_i: for every $w \in W$, if $w' \in B_i(w)$ and $v \in B_i(w')$ then $v \in B_i(w)$
- *Transitivity* of G_i: for every $w \in W$, if $w' \in G_i(w)$ and $v \in G_i(w')$ then $v \in G_i(w)$
- *Euclideanity* of B_i: for every $w \in W$ if $v, v' \in B_i(w)$ then $v' \in B_i(v)$ and $v \in B_i(v')$
- *Euclideanity* of G_i: for every $w \in W$ if $v, v' \in G_i(w)$ then $v' \in G_i(v)$ and $v \in G_i(v')$

Axiom 13 is an axiom of positive introspection for goals similar to axiom 7 for beliefs. Axiom 14 is its negative version (the negative version of axiom 7 is axiom 8). According to axioms 13 and 14, worlds that are compatible with agent *i*'s goals are compatible with agent *i*'s goals from those worlds which are compatible with agent *i*'s beliefs, that is:

- for every $w \in W$ if $w' \in B_i(w)$ then $G_i(w) = G_i(w')$.

Finally, 15 is the *strong realism* axiom studied in [8, 12, 13]. According to this axiom the set of worlds which are compatible with the agent's goals is a subset of the set of worlds which are considered possible by the agent, that is:

- $G_i(w) \subseteq B_i(w)$.

3.1 Validity and Satisfiability

We call $\mathcal{DL}$ the logic axiomatized by the previous axioms 0-15 and rules of inference R1-R5. We call $\mathcal{DL}$ *models* the class of models satisfying all the semantic constraints imposed in the previous section.

We write $\vdash_{\mathcal{DL}} \varphi$ if formula φ is a theorem of $\mathcal{DL}$, viz. if φ is a logical consequence of the set of axioms 0-15 and rules of inference R1-R5.

Moreover, we write $M \models \varphi$ if formula φ is *valid* in the $\mathcal{DL}$ model *M*, viz. $M, w \models \varphi$ for every world *w* in *M*.

We write $\models_{\mathcal{DL}} \varphi$ if formula φ is *valid* in all $\mathcal{DL}$ models, viz. $M, w \models \varphi$ for every $\mathcal{DL}$ model *M* and world *w* in *M*.

Finally, we say that a formula φ is *satisfiable* in a model *M* if there is some world in *M* at which φ is true, viz. there exists a world *w* in *M* such that $M, w \models \varphi$.

Now, we can prove that $\mathcal{DL}$ is *sound* and *complete* with respect to the class of models satisfying all the semantic constraints imposed in the previous section.

Theorem 1. *Soundness and completeness.*
$\vdash_{\mathcal{DL}} \varphi$ *iff* $\models_{\mathcal{DL}} \varphi$

Proof. All axioms and inference rules are in the Sahlqvist class, for which a general algorithm to compute their semantic counterparts exists. Therefore it is a routine to verify that each axiom in 1-15 corresponds to the semantic properties described in the previous section. Furthermore, a general completeness result exists for logics whose axioms are in the Sahlqvist class [14, 15]. Therefore we can conclude that $\mathcal{DL}$ is complete. □

4 Expected Dangers, Dangerous Situations and Expected Attacks

Our general aim is to make clear some categories and concepts which are fundamental for a model of intentional agency with defensive capabilities.

We begin with the assumption that always a defense taken by an agent *i* implies that agent *i* intends either to achieve or to maintain a certain result φ and agent *i* believes that there will be a threat on it.

We suppose that an agent can defend himself from something only if he has predictive capacities. More precisely, if an agent *i* is defending himself from someone then *i* expects that there is an action of another agent *j* which can possibly interfere with the achievement of his goals. In our view a defense always implies an expectation concerning a possible threat or a possible danger, viz. the expectation that an external event can compromise the achievement of our goals.[2]

We can use the formal logic presented in the previous section in order to formalize such an expectation which is always involved in a situation of defense. First of all let us introduce a notational convention.

We write $\bigcirc^n\varphi$ to indicate that the sentence φ is subject to *n* iterations of the modality $\bigcirc$ where *n* can be any number 0,1,2,3,... Therefore 0 is just $\bigcirc^0\varphi$, 1 is $\bigcirc\varphi$, and so on. More formally, $\bigcirc^n\varphi$ can be defined inductively by:

1. $\bigcirc^0\varphi =_{def} \varphi$;
2. $\bigcirc^{n+1}\varphi =_{def} \bigcirc \bigcirc^n \varphi$.

The first concept we are aimed at formalizing is the concept of expected danger.

Definition 1. *Expected danger.*
$ExpDanger(i, j, \alpha, \varphi, \psi, n) =_{def}$
$Bel_i(\psi \rightarrow \bigcirc^n [j : \alpha] \neg\varphi) \wedge Bel_i Goal_i \bigcirc^{n+1} \varphi$

$ExpDanger(i, j, \alpha, \varphi, \psi, n)$ reads: 1) agent *i* expects that, under certain conditions ψ, if *n* steps from now agent *j* does action α then $\neg\varphi$ will hold after α's occurrence and; 2) agent *i* believes that he wants φ to be true *n*+1 steps from now. An alternative reading is: agent *i* expects that, under certain conditions ψ, if *n* steps from now agent *j* does action α then he will interfere negatively with the achievement of *i*'s actual goal that

[2] In our view expectations are a necessary mental ingredients of a BDI like agent. In a previous work [16] we did not introduce expectations as an additional primitive. We preferred to build those mental states on former ingredients (beliefs and goals) in order to have mental states that preserve both properties, epistemic and conative. In the present analysis we make the same kind of assumption by building expectation on the basis of more elementary ingredients (beliefs and goals).

n+1 steps from now φ will be true. Therefore $ExpDanger(i, j, \alpha, \varphi, \psi, n)$ expresses agent *i*'s expectation that the future occurrence of agent *j*'s action α is a danger for him since, given certain conditions ψ, if *n* steps from now *j*'s action α occurs then it will compromise the achievement of his goal that *n*+1 steps from now φ will be true.

For example, $ExpDanger(Mary, Fred, shoot, MaryAlive, inFrontFred, 0)$ means: Mary expects that the occurrence of Fred's action of shooting is a danger for her since if Fred's action of shooting occurs when she is in front of Fred then Fred's action will compromise the achievement of her goal to be alive next.

Starting from the previous definition of expected danger, we can characterize two more specific notions: the notion of expected attack and the notion of expected dangerous situation.

In our vocabulary an agent *i* expects a certain attack if and only if he expects a certain danger and he believes that the danger in question is not simply a potential danger, but it is an actual and effective danger. The concept of expected attack is formalized according to the following definition 2.

Definition 2. *Expected attack.*
$ExpAttack(i, j, \alpha, \varphi, \psi, n) =_{def}$
$ExpDanger(i, j, \alpha, \varphi, \psi, n) \wedge Bel_i \bigcirc^n \langle j : \alpha \rangle \top$

$ExpAttack(i, j, \alpha, \varphi, \psi, n)$ reads: 1) agent *i* expects that, under certain conditions ψ, if *n* steps from now agent *j*'s action α occurs then it will interfere negatively with the achievement of *i*'s actual goal that *n*+1 steps from now φ will be true and; 2) agent *i* believes that *n* steps from now agent *j* will perform action α. Therefore

$ExpAttack(i, j, \alpha, \varphi, \psi, n)$ expresses agent *i*'s expectation that the occurrence of agent *j*'s action α is an actual and effective danger for him (viz. an attack towards him) since *n* steps from now *j* will do action α and, given certain conditions ψ, if *n* steps from now *j* does action α occurs then *j*'s action α will compromise the achievement of *i*'s goal that *n+1* steps from now φ will be true.

For example, $ExpAttack(Bill, thief, forceDoor, moneySafe, nobodyAtHome, 0)$ means: Bill expects that the occurrence of a thief's action of forcing the door of Bill's house is an attack towards him since the thief is going to force the front door and, if the thief's action of forcing the front door occurs when nobody is at home then such an action will compromise the achievement of Bill's goal to keep his money safe.

We suppose that an agent *i* expects to be in a dangerous situation if and only if he expects that if he will be attacked under the actual conditions ψ then one of his goals will be compromised.

The concept of expected dangerous situation is formalized according to the following definition 3.

Definition 3. *Expected dangerous situation.*
$ExpDangerous(i, j, \alpha, \varphi, \psi, n) =_{def}$
$ExpDanger(i, j, \alpha, \varphi, \psi, n) \wedge Bel_i \psi$

$ExpDangerous(i, j, \alpha, \varphi, \psi, n)$ reads: 1) agent *i* expects that, under certain conditions ψ, if *n* steps from now agent *j*'s action α occurs then it will interfere negatively

with the achievement of his actual goal that n+1 steps from now φ will be true and; 2) agent *i* believes ψ to be true. Therefore $ExpDangerous(i, j, \alpha, \varphi, \psi, n)$ expresses agent *i*'s thought that his future-oriented goal that φ will be compromised after every future occurrence of agent *i*'s action α.[3] Going back to one of the previous examples, $ExpDangerous(Bill, thief, forceDoor, moneySafe, nobodyAtHome, 0)$ means: Bill thinks (expects) to be in a dangerous situation since nobody is at home and, if a thief forces the front door when nobody is at home then the thief's action of forcing will compromise the achievement of Bill's goal to keep his money safe.

A further relevant concept of a theory of defense is the concept of expected harm. We suppose that an agent *i* expects a future harm if and only if he expects that he will be attacked by another agent and if he will be attacked under the actual conditions then one of his goals will be compromised.

Definition 4. *Expected harm.*
$ExpHarm(i, j, \alpha, \varphi, \psi, n) =_{def}$
$ExpDanger(i, j, \alpha, \varphi, \psi, n) \wedge Bel_i\psi \wedge Bel_i \bigcirc^n \langle j : \alpha \rangle \top$

We can easily prove that an expected harm implies an expectation of a future frustration of a goal. This is shown in the following theorem of $\mathcal{DL}$.

Theorem 2. $\vdash_{\mathcal{DL}} ExpHarm(i, j, \alpha, \varphi, \psi, n) \rightarrow Bel_i(\bigcirc^{n+1}\neg\varphi \wedge Goal_i \bigcirc^{n+1} \varphi)$

Proof. $ExpHarm(i, j, \alpha, \varphi, \psi, n)$ implies $Bel_i(\psi \rightarrow \bigcirc^n [j : \alpha] \neg\varphi) \wedge Bel_iGoal_i \bigcirc^{n+1} \varphi \wedge Bel_i\psi \wedge Bel_i \bigcirc^n \langle j : \alpha \rangle \top$ (by definitions 1 and 4). Furthermore $Bel_i(\psi \rightarrow \bigcirc^n [j : \alpha] \neg\varphi) \wedge Bel_iGoal_i \bigcirc^{n+1} \varphi \wedge Bel_i\psi \wedge Bel_i \bigcirc^n \langle j : \alpha \rangle \top$ implies $Bel_i \bigcirc^n [j : \alpha] \neg\varphi \wedge Bel_iGoal_i \bigcirc^{n+1} \varphi \wedge Bel_i \bigcirc^n \langle j : \alpha \rangle \top$ (by axiom 5) which in turn implies $Bel_i \bigcirc^n \langle j : \alpha \rangle \neg\varphi \wedge Bel_iGoal_i \bigcirc^{n+1} \varphi$ (by the equivalence $\langle j : \alpha \rangle \top \wedge [j : \alpha] \neg\varphi \leftrightarrow \langle j : \alpha \rangle \varphi$[4]). Finally $Bel_i \bigcirc^n \langle j : \alpha \rangle \neg\varphi \wedge Bel_iGoal_i \bigcirc^{n+1} \varphi$ implies $Bel_i \bigcirc^n \bigcirc\neg\varphi \wedge Bel_iGoal_i \bigcirc^{n+1} \varphi$ (by axiom 4 and rules of inference R2 and R4). □

According to theorem 2 if agent *i* expects a future harm then he believes that he will not achieve something he actually wants.

Agressions as intentional attacks. According to the previous definition 2, agent *i* expects an attack by *j* if and only if *i* expects that agent *j* will perform an action α and the occurrence of α will compromise *i*'s goals. There are more specific types of expected attack which can be reasonably called expected aggressions. Investigating such specific types of expected attack is crucial not only for a better understanding of defense but also for a comprehensive analysis of social interaction.

We suppose that agent *i* expects an aggression by *j* if and only: 1) *i* expects an attack from *j* since he expects that *j* will perform an action α and *i* expects that the occurrence

[3] Indeed $ExpDangerous(i, j, \alpha, \varphi, \psi, n)$ implies $Bel_i(\bigcirc^n \langle j : \alpha \rangle \top \rightarrow \bigcirc^{n+1}\neg\varphi) \wedge Bel_iGoal_i \bigcirc^{n+1} \varphi$.

[4] Verifying that such an equivalence is a theorem of $\mathcal{DL}$ is straightforward (the proof is based on axiom 2 and axiom 4).

of *j*'s action α will compromise *i*'s goal that φ will be true in the future; 2) *i* expects that *j* will perform action α having the intention to do it; 3) according to *i*'s beliefs the possibility that agent *j* already intends to do action α in the future is explained by the fact that *j* believes that *i* wants φ to be true in the future and by the fact that *j* believes that performing α will bring about $\neg\varphi$.

Thus, *i* expects an aggression by *j* if and only if *i* expects a future intentional attack by *j* and *i* thinks that *j*'s attack towards *i* is explained by *j*'s beliefs that doing α will harm *i* (viz. *i* thinks that *j* intends to do α in order to harm *i*).

The complex notion of expected aggression is formalized according the following definition.

Definition 5. *Expected aggression.*
$ExpAggression(i, j, \alpha, \varphi, \psi, n) =_{def}$
$ExpAttack(i, j, \alpha, \varphi, \psi, n) \wedge$
$Bel_i \bigcirc^n Goal_j \langle j : \alpha \rangle \top \wedge$
$Bel_i(Goal_j \bigcirc^n \langle j : \alpha \rangle \top \rightarrow (Bel_j Goal_i \bigcirc^{n+1} \varphi \wedge Bel_j \bigcirc^n [j : \alpha] \neg\varphi)) \wedge$
$\neg Bel_i \neg Goal_j \bigcirc^n \langle j : \alpha \rangle \top$

For example, $ExpAggression(Mary, Fred, shoot, MaryAlive, inFrontFred, 4)$ means:

1) Mary expects an attack from Fred since she expects that 4 steps from now Fred will shoot and she expects that if 4 steps from now Fred shoots and Mary is in front of Fred then Mary will not be alive after Fred's action;
2) Mary expects that 4 steps from now Fred will shoot α having the intention to shoot;
3) according to Mary's beliefs the possibility that Fred already intends to shoot in the future (4 steps from now) is explained by the fact that Fred believes that Mary wants to be alive in the future (5 steps from now) and by the fact that Fred believes that if 4 steps from now he does the action of shooting then Mary will not be alive afterward (viz. if 4 steps from now he does the action of shooting then 5 steps from now Mary will not be alive).

For summarizing, let us make explicit how the five concepts discussed in this section are organized from a logical point of view.

An expected attack is an expected danger
$\vdash_{\mathcal{DL}} ExpAttack(i, j, \alpha, \varphi, \psi, n) \rightarrow ExpDanger(i, j, \alpha, \varphi, \psi, n)$
An expected dangerous situation is an expected danger
$\vdash_{\mathcal{DL}} ExpDangerous(i, j, \alpha, \varphi, \psi, n) \rightarrow ExpDanger(i, j, \alpha, \varphi, \psi, n)$
Expecting a harm is equivalent to expecting both an attack and a dangerous situation
$\vdash_{\mathcal{DL}} ExpHarm(i, j, \alpha, \varphi, \psi, n) \leftrightarrow$
$ExpDangerous(i, j, \alpha, \varphi, \psi, n) \wedge ExpAttack(i, j, \alpha, \varphi, \psi, n)$
An expected aggression is an expected attack
$\vdash_{\mathcal{DL}} ExpAggression(i, j, \alpha, \varphi, \psi, n) \rightarrow ExpAttack(i, j, \alpha, \varphi, \psi, n)$

4.1 Defensive Goals

An expectation of a possible danger is generally responsible for activating and generating defensive goals. In our view a defensive goal of an arbitrary agent *i* should be conceived as a goal of agent *i* which is activated by *i*'s expectation of a possible danger. As we have shown in the previous section, when expecting a danger agent *i* thinks that in a certain situation a certain action of another agent will negatively interfere with the achievement of his goals. Thus, when expecting a danger agent *i* can act in different ways in order to escape the danger: either he can try to block the expected vehicle of attack (*block strategy*) or he can try to get out of the dangerous situation by preventing that the conditions of success of the expected attack are true (*protection strategy*).

The following theorem of $\mathcal{DL}$ shows which kind of defensive goal is activated in the mind of agent *i* when he expects an attack from an agent *j*.

Theorem 3. $\vdash_{\mathcal{DL}} ExpAttack(i, j, \alpha, \varphi, \psi, n) \rightarrow Goal_i \neg\psi$

Proof. $ExpAttack(i, j, \alpha, \varphi, \psi, n)$ implies $Bel_i(\psi \rightarrow \bigcirc^n [j:\alpha]\neg\varphi) \wedge Bel_i Goal_i \bigcirc^{n+1} \varphi \wedge Bel_i \bigcirc^n \langle j:\alpha\rangle \top$ (by definitions 1 and 2) which in turn implies $Bel_i(\psi \rightarrow \bigcirc^n [j:\alpha]\neg\varphi) \wedge Goal_i \bigcirc^{n+1} \varphi \wedge Bel_i \bigcirc^n \langle j:\alpha\rangle \top$ (by axioms 6 and 14). Moreover $Bel_i(\psi \rightarrow \bigcirc^n [j:\alpha]\neg\varphi) \wedge Goal_i \bigcirc^{n+1} \varphi \wedge Bel_i \bigcirc^n \langle j:\alpha\rangle \top$ implies $Bel_i(\bigcirc^n \langle j:\alpha\rangle \top \wedge (\psi \rightarrow \bigcirc^n [j:\alpha]\neg\varphi)) \wedge Goal_i \bigcirc^{n+1} \varphi$ (by standard modal principles) which in turn implies $Bel_i(\psi \rightarrow \bigcirc^n(\langle j:\alpha\rangle \top \wedge [j:\alpha]\neg\varphi)) \wedge Goal_i \bigcirc^{n+1} \varphi$ (by standard modal principles and the equivalence $\bigcirc^n\varphi \wedge \bigcirc^n\psi \leftrightarrow \bigcirc^n(\varphi \wedge \psi)$[5]). Furthermore $Bel_i(\psi \rightarrow \bigcirc^n(\langle j:\alpha\rangle \top \wedge [j:\alpha]\neg\varphi)) \wedge Goal_i \bigcirc^{n+1}\varphi$ implies $Bel_i(\psi \rightarrow \bigcirc^n \langle j:\alpha\rangle \neg\varphi) \wedge Goal_i \bigcirc^{n+1}\varphi$ (by the equivalence $\langle j:\alpha\rangle \top \wedge [j:\alpha]\neg\varphi \leftrightarrow \langle j:\alpha\rangle \varphi$). Finally $Bel_i(\psi \rightarrow \bigcirc^n \langle j:\alpha\rangle \neg\varphi) \wedge Goal_i \bigcirc^{n+1} \varphi$ implies $Bel_i(\bigcirc^n [j:\alpha]\varphi \rightarrow \neg\psi) \wedge Goal_i \bigcirc^n \bigcirc\varphi$ (by the equivalence $\neg \bigcirc^n \varphi \leftrightarrow \bigcirc^n\neg\varphi$[6]) which in turn implies $Goal_i(\bigcirc^n [j:\alpha]\varphi \rightarrow \neg\psi) \wedge Goal_i \bigcirc^n [j:\alpha]\varphi$ (by axioms 4 and 15 and rules of inference R2 and R5). From $Goal_i(\bigcirc^n [j:\alpha]\varphi \rightarrow \neg\psi) \wedge Goal_i \bigcirc^n [j:\alpha]\varphi$ we can infer $Goal_i\neg\psi$ (by axiom 9). □

According to the previous theorem if agent *i* thinks that the occurrence of agent *j*'s action α under the conditions ψ is an actual and effective danger for him (viz. an attack towards him) then he comes to have the defensive goal of getting out of the dangerous situation by preventing that the success conditions ψ of the expected attack are true.[7]

Going back to the examples provided in the previous section, suppose that Bill expects that 4 steps from now he will be attacked by a thief's action of forcing the door of the house since 4 steps from now a thief will force the front door and, if 4 steps from now the thief's action of forcing the front door occurs and nobody is at home, then the thief's action will compromise the achievement of Bill's goal to keep his money safe 5 steps from now: $ExpAttack(Bill, thief, forceDoor, moneySafe, \bigcirc^4 nobodyAtHome, 4)$.

[5] Proving by induction that such an equivalence is a theorem of $\mathcal{DL}$ is straightforward.

[6] By axiom 2 proving that such an equivalence is a theorem of $\mathcal{DL}$ is again an easy task.

[7] With "success conditions" of a vehicle of attack we mean the conditions which ensure that the vehicle of attack will be efficacious and will succeed in compromising the goals of the defender.

Then, according to theorem 2, Bill comes to have the goal that 4 steps from now somebody will be at home:$Goal_{Bill} \bigcirc^4 \neg nobodyAtHome$. In this example Bill decides to defend himself by the thief's attack by preventing that the conditions of success of the thief's attack are true.

The following theorem of $\mathcal{DL}$ is complementary to the previous theorem 2 and shows which kind of defensive goal is activated in the mind of agent *i* when he expects to be in a dangerous situation.

Theorem 4. $\vdash_{\mathcal{DL}} ExpDangerous(i, j, \alpha, \varphi, \psi, n) \rightarrow Goal_i \bigcirc^n [j : \alpha] \bot$

Proof. $ExpDangerous(i, j, \alpha, \varphi, \psi, n)$ implies $Bel_i(\psi \rightarrow \bigcirc^n [j : \alpha] \neg\varphi) \wedge Bel_i$ $Goal_i \bigcirc^{n+1} \varphi \wedge Bel_i\psi$ (by definitions 1 and 3) which in turn implies $Bel_i(\psi \rightarrow \bigcirc^n [j : \alpha] \neg\varphi) \wedge Goal_i \bigcirc^{n+1} \varphi \wedge Bel_i\psi$ (by axioms 6 and 14). Moreover $Bel_i(\psi \rightarrow \bigcirc^n [j : \alpha] \neg\varphi) \wedge Goal_i \bigcirc^{n+1} \varphi \wedge Bel_i\psi$ implies $Goal_i \bigcirc^n [j : \alpha] \neg\varphi \wedge Goal_i \bigcirc^n \bigcirc\varphi$ (by axiom 5 and axiom 15). Finally $Goal_i \bigcirc^n [j : \alpha] \neg\varphi \wedge Goal_i \bigcirc^n \bigcirc\varphi$ implies $Goal_i \bigcirc^n [j : \alpha] \bot$ (by axiom 4 and rules of inference R2 and R5). □

According to the previous theorem, when agent *i* thinks to be in a dangerous situation ψ, since he thinks that if *j* does action α then he will compromise one of his goals, *i* comes to have the defensive goal of trying to block the occurrence of *j*'s action α (viz. the expected vehicle of attack).

Going back to one of the examples provided in the previous section, suppose that Mary thinks to be in a dangerous situation since she believes that one step from now she will be in front of Fred and, if one step from now Fred shoots and she is in front of Fred, then Fred's action will compromise the achievement of her goal to be alive 2 steps from now: $ExpDanger(Mary, Fred, shoot, MaryAlive, \bigcirc inFrontFred, 1)$. Then, according to theorem 3, Mary comes to have the goal of trying to block Fred's action of shooting:$Goal_{Mary} \bigcirc [Fred : shoot] \bot$.

5 For a Specification of Defensive Strategies

In the previous section 4.1 we have analyzed two specific kinds of defensive goals and defensive strategies. We have shown that when expecting a danger an agent can try either to block the expected vehicle of attack (*block strategy*) or to get out of the dangerous situation by preventing that the conditions of success of the expected attack are true (*protection strategy*). The previous two strategies are in our view the most general classes of defensive strategies that an intentional agent can adopt in order to prevent that his goals will be frustrated and compromised. But there are several specifications of these two general defensive strategies and defensive goals. The aim of this section is to provide a very brief overview of such specifications (see also figure 1).

There are two main types of block strategies. We call *objective block* a defensive strategy which consists in blocking the expected vehicle of attack by ensuring that its objective executability preconditions do not hold. For instance, if *i* is defending himself from *j*'s action of shooting *j*, he can try to disarm *j*. Indeed although *j* can intend to shoot *i*, if *j* is not armed then he will not be able to perform the action of shooting *i*.

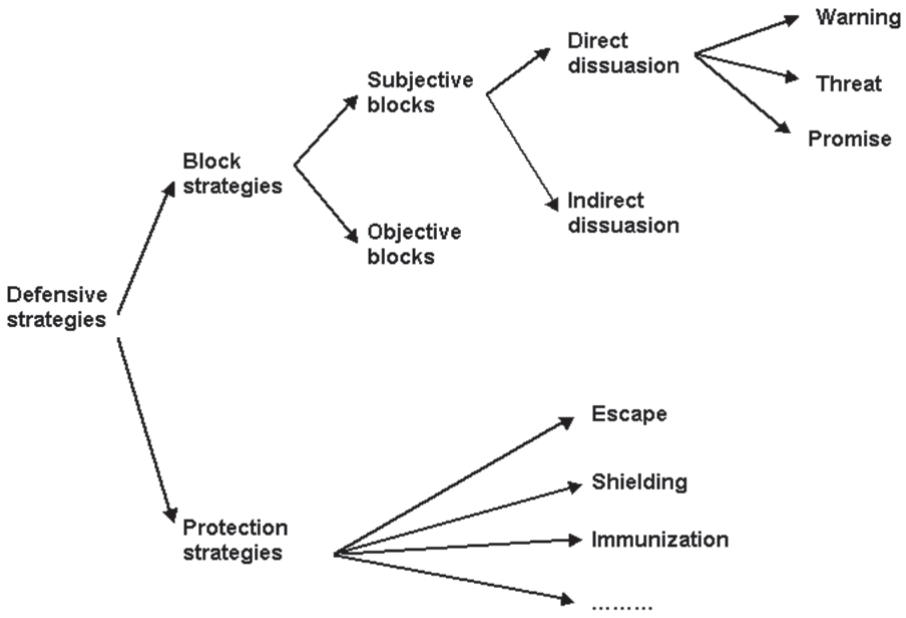

Fig. 1. Typology of defensive strategies

On the contrary, we call *subjective block* a defensive strategy which consists in blocking the expected vehicle of attack by influencing the agent who is supposed to attack.[8] Subjective blocks are suitable defensive strategies only if the expected vehicle of attack is an intentional action of a certain agent. Indeed, when having the goal of influencing agent *j*, agent *i* has the goal of inducing agent *j* to avoid doing such action that according to *i*'s expectations and beliefs is a potential danger for him.

In order to influence agent *j* and to induce him not to do action α, agent *i* should try to modify those beliefs, assumptions, etc... of *j* that (according to *i*) represent *j*'s reasons for doing action α.

There are two general modes of changing the beliefs of another agent *j* in order to induce him not to do a certain action α: we call *indirect dissuasion* one mode and *direct dissuasion* the other mode.

In indirect dissuasion, agent *i* tries to induce *j* not to do a certain action α by changing *j*'s beliefs and undermining *j*'s reasons for doing α without necessarily advancing arguments for not doing α and without necessarily communicating something to *j* in an explicit way.

For example, *i* might try to indirectly dissuade a thief from stealing his car by installing a car alarm. Agent *i* does the action of installing a car alarm (defensive strategy) in order to make *j* believe that stealing the car will be very risky so that *j* will forbear from stealing the car. In this example *i* does not communicate anything to *j*.

[8] See [17] for a theory of social influence.

On the contray, a defensive strategy based on direct dissuasion consists in changing the beliefs of the other agent and undermining his reasons for doing action α by communicating something explicitly to him.

There are several specific types of direct dissuasive strategy such as warnings, threats, promises, etc...[9] For instance, a *threat* of agent *i* to agent *j* should be conceived as a *i*'s act of explicit communication or speech act [20] aimed at informing *j* that: 1) *i* has a conditional intention to perform a certain action β_1 in case *j* will do a certain action β_2; 2) if *i* does β_1 then *j*'s goals will be compromised.[10]

Finally, there are several types of *protection strategies*, viz. defensive strategies aimed at getting out of an expected dangerous situation by preventing that the conditions of success of the expected attack are true. In this work we are not going to deeply analyze them. Let us just note that a protective strategy of escape consists in changing location in order to dodge the expected vehicle of attack whilst a protective strategy of shielding consists in building infrastructures, using artifacts, being protected by law, etc... and generally in building barriers against an expected vehicle of attack.

Generally one would like to distinguish between *preventive defenses* versus *non-preventive defenses*. For the moment, we leave aside this important dimension of defense since in this paper we are mainly interested in the analysis of anticipatory and preventive defenses. Just note that preventive moves are strategic moves that the defender makes before the vehicle of attack starts whilst a non-preventive defense consists either in defending ourselves during the attack (viz. facing the danger) or in defending ourselves after the attack by taking remedies.[11] Moreover, note that the available alternatives for a preventive defense are less than the available alternatives for a non-preventive defense. This statement is validated by observing that agents can block an attack only by preventing it. Therefore we must conclude that a precocious defense is in general more convenient since it offers a wider variety of alternative defensive strategies and moves.

6 Conclusion

We have provided in this paper a general formal analysis of the mental attitudes which are involved in a situation of defense. A preliminary ontology of defensive strategies has been designed. The issue of defense is not totally new in the MAS domain. For instance, the possibility of resolving conflicts through argumentation in negotiation contexts has received a lot of attention in the MAS community.[12] But we think that few efforts have

[9] See also [18] for a theoretical approach to threats in argumentation. For a game theoretical approach to threats see [19].

[10] Note that with explicit communication we mean here the classical gricean conception of meta-level communication [21]. Therefore a threat of agent *i* to agent *j* necessarily implies *i*'s intention to perform a speech act A *in order to* inform *j* that: 1) *i* has a conditional intention to perform a certain action β_1 in case *j* will do a certain action β_2 and if *i* does β_1 then *j*'s goals will be compromised; 2) *i* wants that *j* believes that 1).

[11] This distinction is crucial in medical domain where we can distinguish three macro-phases in the process of medical care: prevention, treatment and rehabilitation.

[12] See [22] for a review of the most important models of argumentation-based negotiation developed in the MAS framework.

been made in order to provide a general and systematic model of defense for intentional agents. As we have shown in this paper in fact, defense by argumentation should be conceived as a particular type of defensive strategy. But there are many other types of defensive moves and defensive strategies which are likewise important and which deserve to be investigated.

We think that with the current formal instruments for modeling intentional action, agency, and in particular with computational models like BDI agents it is possible to arrive to a principled and systematic model of defense. In this work we have tried to build the conceptual basis of such a model.

In our view a model of defense with cognitive and social foundations can be an important reference point not only for modeling security systems but also for modeling important aspects of social interaction such as coordination and negotiation.

References

[1] Castelfranchi, C.: Conflicts ontology. In: Dieng, R., Meller, H. (eds.) Conflicts in Artificial Intelligence, pp. 21–40. Kluwer, Dordrecht (2000)

[2] Burrows, M., Abadi, M., Needham, R.: A logic for authentication. Proceedings of the Royal Society of London 426, 233–271 (1989)

[3] Dixon, C., Gago, M.-C.F., Fisher, M., van der Hoek, W.: Using temporal logics of knowledge in the formal verification of security protocols. In: Proceedings eleventh International Symposium on temporal representation and reasoning. IEEE Computer Society Press, Los Alamitos (2004)

[4] Syverson, P.: Adding time to a logic of authentication. In: Proceedings of the First ACM Conference on Computer and Communications Security. ACM Press, New York (1993)

[5] Glasgow, J., MacEwen, G., Panangaden, P.: A logic to reason about security. ACM Transactions on Computer Systems 10(3), 226–264 (1992)

[6] Emerson, E.A.: Temporal and modal logic. In: van Leeuwen, J. (ed.) Handbook of Theoretical Computer Science, Volume B: Formal Models and Sematics. North-Holland Pub. Co./MIT Press (1990)

[7] Harel, D., Kozen, D., Tiuryn, J.: Dynamic Logic. MIT Press, Cambridge (2000)

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

[9] Hintikka, J.: Knowledge and Belief. Cornell University Press, New York (1962)

[10] Broersen, M.: Modal Action Logics for Reasoning about Reactive Systems. PhD-thesis Vrije Universiteit Amsterdam, Amsterdam (2003)

[11] Castilho, M.A., Gasquet, O., Herzig, A.: Formalizing action and change in modal logic i: the frame problem. Journal of Logic and Computation 9(5), 701–735 (1999)

[12] Herzig, A., Longin, D.: C&L intention revisited. In: Dubois, D., Welty, C., Williams, M.A. (eds.) Proceedings 9th Int. Conf. on Principles on Principles of Knowledge Representation and Reasoning (KR 2004), pp. 527–535. AAAI Press, Menlo Park (2004)

[13] Miller, K., Sandu, G.: Weak commitments. In: Holmstron-Hintikka, G., Tuomela, R. (eds.) Contemporary Action Theory. Social Action, vol. 2. Kluwer Academic Publishers, Dordrecht (1997)

[14] Blackburn, P., de Rijke, M., Venema, Y.: Modal Logic. Cambridge University Press, Cambridge (2001)

[15] Sahlqvist, H.: Completeness and correspondence in the first and second order semantics for modal logics. In: Proceedings 3rd Scandinavian Logic Symposium 1973. Studies in Logic, vol. 82 (1975)

[16] Castelfranchi, C., Lorini, E.: Cognitive anatomy and functions of expectations. In: Schmalhofer, F., Young, R.M., Katz, G. (eds.) Proceedings European Cognitive Science Conference 2003 (EuroCogSci 2003). Lawrence Erlbaum, Mahwah (2003)
[17] Conte, R., Castelfranchi, C.: Cognitive and social action. London University College of London Press, London (1995)
[18] Walton, D.: Plausible Argument in Everyday Conversation. State University of New York Press, Albany (1992)
[19] Schelling, T.C.: The strategy of conflict. Harvard University Press, Cambridge (1960)
[20] Searle, J.: Speech acts: An essay in the philosophy of language. Cambridge University Press, Cambridge (1969)
[21] Grice, H.P.: Study in the way of words. Harvard University Press, Cambridge (1989)
[22] Rahwan, I., Ramchurn, S.D., Jennings, N.R., McBurney, P., Parsons, S., Sonenberg, L.: Argumentation-based negotiation. The Knowledge Engineering Review 18(4), 343–375 (2003)