Reasons: Belief Support and Goal Dynamics

Cristiano Castelfranchi
Nat. Research Council - Inst. of Psychology
e-mail: cris@pscs2.irmkant.rm.cnr.it
IP-CNR, v. Marx 15, 00137 Roma. Italy

Abstract

The paper is devoted to the structural relation between beliefs and goals. I discuss its importance in modelling cognitive agents; its origin in cognitive processing; its stucture (*belief structure* relative to a goal); its crucial role in rationality, mediating between "epistemic" and "pragmatic" rationality; its role in goal Dynamics. I stress the crucial contribution of the supporting beliefs to the *Processing of goals*; to the*Revision of goals* (or *Dynamics* in a narrow sense), i.e. the change of goals either on the basis of the change of a dynamic external environment, or of internal cycles of the agent; and to the *Typology of goals*, that may be partially characterized just on the basis of their typical *belief structure.* In particular, I will analyse in this paper the role of beliefs in the *Processing of goals*, from their firing to their satisfaction or giving up: how beliefs determine such a process step by step. The paper will not give a complete or formal account of any of these aspects. It is more an exploratory paper, which tries to identify basic ontological categories and principles, and fruitful directions of analysis for modelling the relation between beliefs and goals.

1 Premise

1.1 Why the belief-goal bridge is so important

There is a very crucial and promising area that challenges AI studies in the nearest future. Namely, it is the issue of cognitive based goal Dynamics, or better, of *the architectural relationships between beliefs and goals* (wishes, intentions, preferences, and so on). This topic is extremely crucial for several reasons.

The kernel

The belief-goal relation constitutes the kernel of the Cognitive agent architecture.

A Cognitive agent, as a matter of fact, *is an agent who founds his decisions, intentions and actions on his beliefs.*

This bridge connects knowledge, and abstract mechanisms of "cognitive processing", to action and then to adaptation.

> 1a) Through this bridge goals impact on and influence the processes of knowledge acquisition, retrieval, maintenance, and introduce some relevant bias [5].
>
> 1b) Through this bridge beliefs determine goals life: their activation, choice, generation and planning, dropping, etc.

The second phenomenon will be the main topic of the paper.

How epistemic rationality impacts on pragmatic rationality

The *belief structure* supporting goals, i.e. Reasons to Do (or not to Do) explain also which is the relation between "epistemic rationality" (laws of soundness, consistency, rational credibility, etc. in knowledge management) and "pragmatic rationality" (rationality governing decisions and actions): since goals are supported by beliefs and processed on the basis of beliefs, then *rationality in believing contributes to rationality in behaving.*

Rational beliefs are a necessary condition for rational behaviour, since irrational beliefs are a sufficient condition for irrational behaviour (from an observer's point of view).

> a) If I prefer the most reliable source of knowledge or the most credible belief when revising my beliefs base, my trust in and investment on a certain goal on the basis of these beliefs has greater probability of success. Obviously, if I choose my beliefs either randomly or preferring the less supported, or the less believable ones, my plans and actions based on such beliefs could be very risky.
>
> b) A complex action or plan, based on inconsistent beliefs, is necessarily incoherent and self-defeating. If agents do not care about a sufficient consistency of their beliefs, they will not construct rational plans and take rational decisions.

Social Perspective

Many developing AI areas have touched this topic, just tasting some of its aspects. In fact, beliefs based goal Dynamics is a fundamental step for the theories of argumentation, negotiation, persuasion, explanation, as well as for language generation, cooperation among autonomous agents, conversation, or plan and intention recognition. Why is this topic so crucial, and why is it an unavoidable passage for all these research lines? The answer does not lie only in the previous definition of Cognitive agent. The main reason is what I would call the "Autonomous Cognitive Agent Postulate":

> *(1) It is impossible to directly modify the goals (and then the intentions and actions) of an Autonomous Cognitive Agent. In order to influence him (i.e. to modify his goals), another agent should modify his beliefs supporting those goals.* [5]

Thus, the control over beliefs becomes a filter, an additional control over the adoption of goals.

1.2 Aims

Of course, the topic of this work is strongly related to the theory of action, the epistemic logic, the BDI architectures. Some elements of the goal Dynamics theory have already been identified, and there are some proposals for a formal treatment of these elements. However, in my view, the logic instrument is anticipating the identification of the aspects of the cognitive processes we want to model. Sometimes, it seems that the instrumental apparatus distorts and limits the preliminary clarification of the crucial notions. My methodological claim is that we need an *ontology* of goals, intentions and commitments [6], etc. (which, in some sense, is even preliminary to any logic), and that the current ontological discussion is insufficient and a bit biased by logic. Suffice to mention the strange attempt to arrive at a theory of the relationships between intentions and beliefs, and of the intentions Dynamics, without a theory of the relationships between goals and beliefs [19], and with a very unsatisfactory notion of "goal".

As I said, part of this theory of belief-goal relations has been already proposed in AI logic for action or in BDI models ([9] [10] [4] [21] [2] [24]), and also the idea of a "*belief structure*" supporting goals -crucial in the process of goal adoption and persuasion- is already existing [22]. Nevertheless, as far as I know, no specific study exists of such structure, of the basic kind of support and of the multiple roles of such structure.

In general, what Bratman [3] calls "coherence" and Cohen & Levesque's "rational equilibrium" between the agent's intentions and beliefs [10], is reduced only to the fact that the agent selects and adopts those intentions that he believes to be achievable. More precisely:

- In current BDI models, beliefs are of course crucial for the adoption or the abandoning of intentions, but their role seems quite limited: during the processing the belief component is not consulted at each step (consider for ex. Georgeff-Rao architecture): some crucial steps, like planning, are not based on beliefs (means-end and causal relations). Only in Bratman, Israel and Pollack architecture [4], beliefs enter all the components of the architecture, determining activation, deliberation, planning, etc. In some sense, I will just make explicit such a role of beliefs in the process, adding also the idea of their “supporting” role, and of their effect on the “quality” of the goal.

- In those models there is no clear distinction among:

a) the *Processing of goals*, from their firing to their satisfaction or abandon: how beliefs determine such a process step by step;

b) the *Dynamics or Revision of goals*, i.e. the change of goals (“motivations”, “preferences”, “desires”, depending on the terminology of different authors) on the basis the change of a dynamic external environment, or internal cycles of the agent; and

c) the *Typology of goals*, that may be partially characterized just on the basis of their typical *belief structure.*

Of course, there are relations among these different aspects of goal theory in which *belief structure* is relevant. Normally, the processing of a goal from its firing to its satisfaction is intertwined with the Dynamics of goals (changing goal, or the activation of other goals, etc.). But in principle one could just follow the processing of a single goal from its conception to its pursuit, showing the role of beliefs in this flow. Also the differences among kinds of goals (like “intentions” Vs “desires”, or “expectations” Vs “renounces”, etc.) are frequently related to different steps of the goal processing.

The paper will not give a complete or formal account of any of these aspects. It is more an exploratory paper, which tries to identify fruitful distinctions and directions of analysis, and reach an overall view of the role and importance of cognitive Reasons for supporting goals. A general theory of this relation is needed, that should include, in my view, four claims about the role of beliefs relative to goals’ life:

- beliefs support goals (they become their Reasons);
- beliefs determine goal processing;
- beliefs determine goal dynamics;
- beliefs determine goal kinds.

However, in this paper I intend to support only the two first claims: to describe how beliefs “support” goals, and how goals are activated, instantiated, generated, chosen, (adopted from the exterior), abandoned, put aside, just on the basis of their related beliefs.

2 Beliefs support goals

I will base my analysis on the following postulate:
"Postulate of Cognitive Regulation of Action"

> *(2) In a cognitive agent goals should be supported and justified by the agent's beliefs (Reasons).We can not activate, maintain, decide about, prefer, plan for, pursue, goals which are not grounded on pertinent beliefs.*
>
> Corollary 1: *In each phase of their processing, goals are supported by specific beliefs*, that determine the new "quality" of the goal (e.g. from wishes to intentions).
>
> Corollary 2: *The destiny of a goal after its invalidation strictly depends on the reasons of its invalidation (specific invalidated belief), on the kind of goal, and on its processing stage.*

There are *many reasons* for dropping a goal and there are *many* consequent *destinies* of the dropped goal.

Since process and flow depend on beliefs and beliefs are modifiable by the other agents, through communication or perception, at the social level, from Corollary-1 it follows that:

> *(3) At any level of the goal processing, during any phase, the autonomous agent is exposed to some external influencing action aimed at changing his goals.*

2.1 Support relations: their meaning

We talk about "support" relations, because cognitive items hold thanks to such relations. There is a specific structure of beliefs (with their relative structure of justifications and supports) necessary to maintain and justify a certain goal. To "support" means that *without such beliefs the related goal would be dropped from its current state* (not necessarily discarded): it will change its nature and status.

> > To maintain a goal (in a given status) you should maintain its supporting beliefs.
>
> > Revising these beliefs involves revising their goal.

Just to give a trivial example, suppose you are taking a train to go to Naples, and you discover that this train is not going to Naples. You will abandon that train (the goal to take that train) which was based on the belief that to take that train was useful to go to Naples. Suppose that in the same situation you discover that there is another train for Naples that arrives earlier. Again you could abandon the train, as far as your goal was supported by the belief that this was your best option.

A *Directly Supporting belief* is a belief that directly supports a goal (determining its status and nature). I think that Directly Supporting beliefs are beliefs "about that goal". This means either a belief about the content of the goal (be it a world state or an action), e.g.: (GOAL x p) & (BEL x p); or a meta-belief about the goal it self, as a mental object, e.g.: (BEL x (GOAL x p)).

An *Indirectly Supporting belief* is a belief which "supports" a Directly or Indirectly Supporting belief.

Not all beliefs "about" the goal or its content are supporting beliefs of goal p. Many of them are unimportant. Which beliefs relatives to p (or to goal p) are supporting beliefs of goal p?

The belief-goal supporting relation is not only a logical relation; it is also a "historical" one: only the beliefs that entered in the activation, or generation, or adoption, or choice, or planning, or persecution, etc. of goal p are later "supporting" it.

For example the belief "Bill wants me to go to Naples" (BEL x (GOAL y (GOAL x p))) can be completely irrelevant for x to have or pursue goal p. If x has his own independent reasons for goal p and does not care about y's desires, such a belief is not supporting x's goal. If, on the contrary, this belief was the reason why x adopted such a goal (Adoption belief) -to meet y's expectations- then it is a "supporting" belief of goal p: x will give up such a goal if (he believes that) y does not want it any longer.

2.2 Support relations: their origin

Let now examine a quite important cognitive property:

> *(4) Any process of knowledge acquisition, generation or elaboration does not only generate the output knowledge item; it also generates a "relation" that supports and integrates such a new item.*

In other words, we maintain in our memory a trace of the derivation of the cognitive item: its story (I will call this: *trace hypothesis*). By "trace" I do not mean the detailed trace of all the process, step by step. Frequently enough cognitive processes are silent and uninspectable: we do not perceive them; we ignore them. A trace means a link, a relation between the *source* of the knowledge acquisition or elaboration, and the result. This link maintains also the kind of derivation: perception, communication, reasoning.

Consider for example inferences.

Inferences

By "inference" I mean *the process by which a cognitive system is able to generate, internally, new pieces of knowledge from already existing pieces of knowledge explicitly represented.* It is a knowledge acquisition process.

There should be then some *rule* or *procedure* that uses some existing beliefs as input (*premises*) -independent of their origin (be it perceptual, or from communication, or again inferential)- and produces a new belief (*conclusion*).

Ex. John killed Mary ⇒ Mary has died

John is at the Workshop + the Workshop is in Naples ⇒ John is in Naples (See fig.1 A).

Thus, inference is one of the *sources* of the system's knowledge. This sources is normally less reliable than perception [23], but I will not discuss here credibility principles, and in particular degrees of credibility for inferred beliefs.

Let me now generalize this notion of a generative cognitive processing. It is clear that it applies only to propositions. However propositions may be object of different kinds of "mental attitudes", not only knowledge and beliefs, but also for example goals. So we could consider as "inference rules" also those *rule and procedures that derive new goals from preexisting active goals*: like in planning.

Practical syllogism, which is also the basic planning rule, may be considered an "inferential" device (fig. 1 B):

To infer a goal necessarily one of the premises should be a goal (one cannot generate the goal of "killing Mary" just from the belief that "killing Mary" implies that "Mary dies", without the goal that "Mary dies").

To infer a goal necessarily one of the premises should be a belief/knowledge (one cannot generate the goal of "killing Mary" just from the goal that "Mary dies" without the belief that "killing Mary" implies that "Mary dies").

Thus, given our trace-hypothesis *(4)*, arrows in fig 1 A and B, not only represent a process, a generation in time, they also represent a static, *structural relation that remains (as a result of that process) among cognitive items.*

This trace and relation may have a merely procedural or structural nature (*implicit supports*); but it could also be made explicit through a declarative link: a proposition, a belief that explicitly states such a relation. I will call *Reasons* these *explicit support* s (examples: "r because p & q", "since p & q, then r", "the fact that p & q justifies why I believe r", etc.). So, "Reasons" are meta-cognitive objects, based on some form of meta- memory [8].

Processes of knowledge acquisition and generation generate at the same time knowledge structure (network) or "relational knowledge": Reasons.

This is true not only for inferences, but in general. Knowledge items remain related to their *source*: "I saw that p"; "I think that p, because..."; "the TV said that p", etc. Thus, there is a special relation between the belief that p, and the belief "I saw that p" or "the TV sad that p".

In this view, the *belief about the source* together with the *belief about the reliability of the source*, are "reasons" for believing, and support the adopted belief: if this belief is revised one should revise one of those reasons [12], and viceversa.

Consequences of trace theory are the following ones:

- Items are *integrated in cognitive nets*: one cannot eliminate or insert a new item of knowledge, without dealing with its supports and relations. And one would not be able to revise his knowledge and to maintain its coherence without such links. Changes are never merely local.

- Part of the difference between various "mental attitudes" (like: belief, knowledge, opinion, prediction, etc.) is to be traced back to the "story" and the support of the proposition: its "Reasons".

- We maintain in our mind both: *Reasons to believe*, and *Reasons to Do.* We have to have "reasons" both for believing and for "goaling". We cannot do this arbitrarily. This is the *common feature of both faces of our "rationality"*: belief rationality (epistemic), goal rationality (pragmatic).

- "Reasons" give the agent the possibility to *justify* and *explain* (to itself and to the others) its actions, being in this way a major aspect of its *rationality* and of its consciousness.

I will not discuss in the paper either the *Reasons to believe* or the belief-belief support relations. The latter is also a quite well studied problem in AI (truth maintenance systems; belief revision and updating; argumentation).

The view I illustrated is close to Pollock's model [20]. However, the latter is just about beliefs; moreover he calls "reasons" the structural relations among beliefs, not the explicit beliefs about these links; finally he does not claim that there is a memory of the origin and source of the piece of knowledge.

This view is also strongly related to the so called "foundations approach" (as opposed to the "coherence approach") in belief revision ([17] [15] [16] [13] [14]). In particular, in Doyle's approach also "desires" are admitted to have "reasons" and maintenance of these reasons (however the model has been really developed for beliefs). My view differs from the foundations approach for the following aspects: the inclusion of the source among the reasons; the memory of the source; the idea that inference are a kind of source; and the fact that I consider only the *invalidation*, the retraction of a reason -not simply its forgetting- as a reason for the revision of the supported items (and viceversa).

On the contrary, I do not think that the current view of *relations among goal* i.e. how goals support other goals, is satisfying. Neither the inconsistency-conflict relation, nor the inferential or planning (means-end) relation are very well analysed. For ex., Cohen and Levesque's [9] notion of "relativized" goal seems too strong and confused: it mixes up instrumental goals and other completely different relations. It seems that any "reasons" for dropping a goal, any beliefs supporting a goal can be expressed as a clause which the goal is relativized to. As Konolige and Pollack [19] claimed we need some specific definition of the means-end relationship.

2.3 Supporting beliefs and their structure

Not only beliefs support beliefs, and goals support goals (top goals motivate sub-goals; sub-goals make top goals achievable): *goals are also supported by many beliefs*. Not only by the already mentioned belief of means-end relation which is also the belief of the planning rule (or practical syllogism) and the scheleton of every plan. Any active goal has a specific *belief structure* that support it and gives its Reasons; any kind of goal has its typical frame of beliefs. The kind of beliefs depends on the kind of goal and/or the level of processing.

The following is just a list of possible beliefs that support goals:

- *Triggering beliefs*: beliefs that reactively activate goals on the basis of a pre-established association. ex: belief: Fire alarm ⇒ goal: to escape;
- *Conditional beliefs*: beliefs that activate a goal on the basis of the conditional nature of the goal in itself; ex: belief: It is Sunday ⇒ goal: to go to the mass if/when it is Sunday;
- *Adoption beliefs*: ex.: belief: She wants/needs that I do a ⇒ goal: to do a;
- *Satisfaction belief*: I have the goal that p at time j, and at time j (or >j), I assume that p is or was true at time j;
- *Impossibility belief*: it is impossible that p at time j, or p is never possible;
- *Preference belief:* I believe that goal G1 is better than goal G2;
- *Urgency belief*: I believe that G1 is more urgent than G2;
- *Compatibility belief*: I believe that G1 is compatible with G2;
- *Cost belief*: I believe that the cost of a given action or plan is such and such (I know how much I should spend to pursue G1);
- *Value belief*: I believe that the Value of G1 is such and such;

- *Know-how beliefs*: I believe that I know how to reach the goal (some plan that achieves it);
- *Means-End beliefs*: I believe that G2 is useful for G1: if achieved it will cause or allow the achievement of G1;
- *Cando beliefs*: I believe that I have in my action repertoire the actions that are sufficient to reach the goal (given my know-how relative to how the goal can be reached);
- *Condition belief*: I believe that external conditions and resources necessary for the successful execution of the actions, hold.

This list is neither complete, nor ordered, nor well defined. Besides, one may have "complex beliefs" or "attitudes" based just on the combinations of these beliefs.

Why I say that there is a "structure" of beliefs around a goal? In fact, it is not just a list, for three reasons:

- first, there is a sort of typical "frame" around any kind of goal (ex.: an "achievement" goal, [10]), specifying the kinds of belief one have to have for such a goal;
- second, these beliefs have their own supports, thus in fact there is a small belief network: to demolish a supporting belief, you have to normally "revise" its related part of the belief network;
- third and more relevant, *there are relations also among some of these beliefs supporting goals.* For example, the *Impossibility belief* may be derived by some intrinsic impossibility of the goal p ("to be married and to be free") but it could be also derived by a negative *CanDo belief* or *Condition belief.* The *Urgency belief* is based on a belief about the deadline of G1, a belief about the deadline of G2 (which are some of the beliefs supporting respectively G1 and G2), and a belief about the ordering of the two deadlines.

In other terms: beliefs supporting goals, may be in structural relations among each other; they may support one the other or be components of complex beliefs.

3 The Role of beliefs in goal Processing

(5) Beliefs determine goal Processing

In this section I will assume a very simplified model of goal Processing -inspired both by psychological models [18] and by BDI architectures (mainly Georgeff)- and I try to show how the presence or the absence of a specific belief will determine the flow of the goal in one direction or another, its moving from one step to the following one: *beliefs are test conditions in this flow.*

I supposed 8 goals state (Sleeping; Active that includes Fired, Chosen, Planned, Intended, and Executive; Waiting; Dead) and 6 transformations (Activation; Deactivation; Dropping; Decision; Planning; Commitment; Execution). Fig 2

In the BDI architectures all goals "originate" from desires or wishes, but this is misleading. In my view, only a sub-set of Active goals are desires [11]. We cannot consider as "desires" those goals or intentions that derive from obligations, duty or coercion! In general, all plans and instrumental goals (sub-goals) generated to satisfy a desire are not necessarily desires. I admit *several different goal sources* or origins: goal Activation from long term memory based on beliefs [1] [7]; emotional activation of "impulsive goals"; physiological activation of goals (ex. hunger); goal-Adoption from external requests, norms, commands, etc.[11]; sub-goal generation in planning. I will consider in this paper only one source: the belief based activation of sleeping goals from long term memory (we may mainly consider these fired goals as "desires" in BDI sense).

Consider I have a couple of *active goals* (goals I'm considering in order to decide *if* and *which* to pursue) (different authors call these: *wishes, desires, preferences*): G4 "to prepare the lunch" and G7 "to go to the mass". They were sleeping, but both of them were activated by certain beliefs ("It is Sunday"; "It is 10 a.m."). They both are possible and not already achieved. To be considered as alternative options two (or more) goals should be believed as "incompatible"; the chosen goals have to be "compatible". For a goal to be chosen, we have to believe that it is preferable to the other incompatible goals. So in order to access the next processing stage the goal G4 should have the following beliefs (test could be either in series or in parallel, this is not in the model):

Compatibility belief, or, if incompatible with G7, a *Preference belief* (G4 is better than G7) possibly rationally derived from some *Value beliefs* relative to both G4 and G7; or an *Urgency belief.*

On such a basis, suppose I decide to pursue G4 and not G7 which lacks of some of these beliefs. In order to pass to the next step and produce an intention and an executive goal, I have to generate some plan or sub-goal for G4: G4a or G4b. To do this, other beliefs are necessary: *Means-End beliefs.* Again I have to choose, and I need some beliefs: for example, the *Cost belief* (how much I should spend pursuing G4a or G4b) and the consequent *Convenience belief* (the Value of G4 exceeds its costs, and the plan G4a is better than G4b). One could say that this belief is already necessary to choose between G4 and G7, but this is not necessarily true: this is only the "economic rational" strategy; one can have different heuristics. We need also *Know How beliefs* and *CanDo beliefs.* Later, to arrive to the status of "execution" our goal (now is an intention that relates the instrumental goal G4a to its end-goal G4) should have also its *Conditions beliefs.*

Thus, *for a goal to arrive to the last stage (pursuing/execution), all the supporting beliefs are necessary.* If we look at the state-goal (the result to be reached) its last stage is "pursuing": to be pursued one should have planned it, should have the actions necessary for that plan, their conditions and resources. If we look at the "action-goal" (to do a certain action is a goal), actions will be "executed" and

requirements are the same: to be chosen (choosing a plan), to be able, to have conditions and resources. But also the beliefs of the previous stages have to remain true.

If a belief that determined the progression of a goal to the next stage, is invalidated, the goal is eliminated by that stage. But it does not necessarily disappear (Dead goals): it can regress in the process, can be put in some waiting room, or be postponed, can sleep again, or be completely abandoned.

It is quite important to notice that Chosen goals are a sub-set of Option goals, that are a subset of Fired goals, that (ignoring other sources) are a subset of Sleeping goals; but, planned instrumental goals are not a subset of the original set of goals (Sleeping, Fired, "desires" or whatever): they are *a new set of active goals generated by planning* (this is why in the figure there is a sub-section in the process). Again in this set there is a selection process: executed intentions are a subset of executable intentions, that are a subset of chosen intentions, that are a subset of optional plans.

One could say that *intentions inherit their supporting belief structure* both *along the process* from previous goal-steps, when a goal inheriting all this beliefs and adding other "becomes" an intention; and *hierarchically*, from the class of intention which has its typical structure.

4 Conclusions

I hope that at the end of the paper the reader be more aware of the importance of beliefs-goals structural relation (*Reasons for goals*) in modelling cognitive agents, its origin in cognitive processing, its structure, its crucial role in rationality, mediating between "epistemic" and "pragmatic" rationality, its crucial role in goal Dynamics (in a broad sense). I analysed the role of beliefs in the *Processing of goals*, from their firing to their satisfaction or abandon: arguing how beliefs should determine step by step such a process. I suggest some crucial issues (to be addressed in future work) relative to the *Dynamics or Revision of goals*, and to the *Typology of goals*, that may be partially characterized just on the basis of their typical *belief structure*. I know that all these analyses were quite disappointing, because of their lack of completeness and formalization. However, I hope that the reader could recognize that such a preliminary "ontological" work is both unavoidable and useful for modelling cognitive agents. Notice that no "quantities" were introduced in the discussion, bur clearly enough many of these processes (activation, preferences, urgency, possibility, inferences, support, etc.) need also a quantitative treatment.

References

[1] Basso, A., Mondada, F., Castelfranchi, C. 1993. Reactive Goal Activation in Intelligent Autonomous Agent Architecture. In Proceedings of AIA'93 - First

International Round-Table on "Abstract Intelligent Agent", ENEA, Roma, January 25-27.

[2] Bell, J. 1995. Changing Attitudes. In M. Woolridge and N. Jennings (eds.), *Intelligent Agents: Theories, Architectures and Languages.* LNAI 890, Springer Verlag.

[3] Bratman, M.E., 1990. What is an Intention? In *Intentions in Communication.* P.R. Cohen, J. Morgan, M.A. Pollack (eds),pp.15-32. Cambridge, Mass.: MIT Press.

[4] Bratman,M.E., Israel, D.J., Pollack, M.E. 1988. Plans and resource-bounded practical reasoning. *Computational Intelligence* 4: 349-55.

[5] Castelfranchi, C., 1995. Guarantees for autonomy in cognitive agent architecture. In M. Woolridge and N. Jennings (eds.), *Intelligent Agents: Theories, Architectures and Languages.* LNAI 890, pp.56-70, Springer Verlag.

[6] Castelfranchi, C., 1995. Commitments: From Individual Intentions to Groups and Organizations. In V. Lesser (ed.) ICMAS-95: Proceedings of the First International Conference on Multi-Agent Systems. AAAI-MIT Press, pp. 41-8.

[7] Castelfranchi, C., D'Aloisi, D., e Giacomelli, F. 1995. A framework for dealing with belief-goal dynamics. In M. Gori and G. Soda (eds.), *Topics in Artificial Intelligence* (pp. 237-242). Berlin: Springer-Verlag.

[8] Cavenaugh J.C. & Perlmutter, M., 1982, Metamemory: A Critical Examination, "Child Development", 53, 11-28.

[9] Cohen, P.R., Levesque, H.J. 1990a. Intention is choice with commitment. *Artificial Intelligence*, 42, pp. 213-261.

[10] Cohen, P. R. & H.J. Levesque 1990b. Rational Interaction as the Basis for Communication. In *Intentions in Communication.* P.R Cohen, J. Morgan, M.A. Pollack (eds), 33-71. Cambridge, Mass.: MIT Press.

[11] Conte, R. & C. Castelfranchi. 1995. *Cognitive and Social Action*, London: UCL Press.

[12] Dragoni,A.F., 1992. A Model for Belief Revision in a Multi-Agent Environment. In *Decentralized AI - 3*, Y. Demazeau, E. Werner (eds), 215-31. Amsterdam: Elsevier.

[13] Doyle, J.,1979. A truth maintenance system. *Artificial Intelligence* 12(2): 231-72.

[14] Doyle, J., 1992. Reason maintenance and belief revision: Foundations versus coherence theories. In P. Gardenfors (ed.) *Belief Revision.* Cambridge University Press, Cambridge, UK, pp. 29-51.

[15] Gardenfors, P., 1989.The Dynamics of Belief Systems: Foundations vs Coherence Theories. *Revue Internationale de Philosophie*, 172: 26-46.

[16] Gardenfors, P., 1992. Belief revision: An introduction. In P. Gardenfors (ed.) *Belief Revision.* Cambridge University Press, Cambridge, UK, pp. 1-28.

[17] Harman, G., 1986.*Change in View: Principles of Reasoning,* MIT Press, Cambridge, MA.

[18] Heckhausen, H., and Kuhl, J. 1985. From wishes to actions: The dead ends and short cuts on the long way to action. In M. Frese & J. Sabini (eds.), *Goal-directed Behavior: The concept of action in psychology.* Erlbaum. N.J.

[19] Konolige, K., Pollack, M.E., 1993. *A Representationalist Theory of intention* 13 IJCAI'93, Chambery, France, sept. 1993, pp.390-95

[20] Pollock, J.L., 1989. *How to Build a Person: A Prolegomenon.* MIT Press. Cambridge, MA.

[21] Rao A.S., Georgeff, M.P., 1991. *Modelling rational agents within a BDI-architecture*, KR91, Cambridge, MA, pp.473-484

[22] Sycara, K., 1991. Pursuing persuasive argumentation. In: *Symposium on Argumentation and Belief.* AAAI Sping Symposium Series. Stanford University.

[23] van Linder, B., van der Hoek, W., Meyer, J-J. Ch., 1995. Seeing is Believing. And so are Hearing and Jumping. In M. Gori and G. Soda (eds.), *Topics in Artificial Intelligence.* LNAI 992, Springer Verlag, pp. 402- 13.

[24] van Linder, B., van der Hoek, W., Meyer, J-J. Ch., 1995. How to Motivate Your Agents. In In M. Woolridge et al. (eds.), *Intelligent Agents 2: Theories, Architectures and Languages.* LNAI, Springer Verlag.

Acknowledgements

A preliminary and extended version of this paper has been presented at the *2st Round Table on Abstract Intelligent agent - AIA'94* 20, Roma, ENEA, February 1994. Current developments have been supported by the European ESPRIT3 Working Group *"MODEL AGE"* and by the Italian Progetto Coordinato on "Explanation in Intelligent Systems" CNR 1995-96 (Resp. L.Console). I would like to thank Maria Miceli for her precious comments and help.