Reasons to Believe: Cognitive Models of Beliefs Change

Cristiano Castelfranchi

Th□ - Theoretical Psychology group
Institute for Cognitive Sciences and Technologies, National Research Council (ISTC-CNR)
Viale Marx 15, 00137 Roma, Italy – c.castelfranchi@istc.cnr.it

1. Belief Adoption and Its Constraints

It is a fact of life that we cannot believe everything we observe or that we are told. We accept a given information or datum as a belief on the basis of our previous beliefs, of its evidences, supports and sources, and of others psychological factors. Here I will sketch some crucial points of these cognitive mechanisms.

Our knowledge base is not a file where one can introduce new data or eliminate a file-card without altering the other data. Our beliefs are integrated, interconnected and mutually supported: to drop a belief or to add a new one entails checking its coherence with other beliefs and revising previous knowledge. The belief-belief coherence and support is quite a well studied problem in philosophy and AI (truth maintenance systems; belief revision and updating; argumentation) and in some cognitive agent architectures. There are in fact two schools in belief revision (Harman 1986; Gärdenfors 1988; Doyle 1992): the "foundations approach" stressing the importance of supports and justifications of beliefs, and the "coherence approach" modelling logical compatibility and coherence. However, we agree with Doyle (1992) that there is no incompatibility between the two models, and that beliefs must be both relatively coherent and justified.

1.1 To Accept and to Reject; Storing vs. Believing

It has been recently remarked (Castelfranchi 1997; Paglieri 2004) that the meaning of 'revising' and of 'rejecting' a belief is quite obscure. One either believes something or she does not believe it. What does it means to 'reject' a belief or an information (Cantwell 1996) (which is not the same)? In our view in many approaches there is a dangerous confusion between memory and knowledge, between storing an information and accepting/believing it. In several models, the belief base is imagined as a memory store plus the coherence constraint. When a new belief arrives it is either coherent -and it is accepted (added)- or it is in conflict with other beliefs. In that case, after a given process, the new belief is either rejected (not stored) or the old knowledge is revised in order to be compatible with the new information. In this simplistic scheme, to ‘accept’ means both to store and to believe, while to ‘reject’ means not to believe and even not to store the ‘information’. In order to avoid such a confusion and to clarify our interpretation of accepting and rejecting, we will assume *a strong independence between memory and knowledge* (or better beliefs). Two points appear rather obvious (and yet overlooked so far in formal models):

a) one can remember something she does not believe (and remember she does not believe it);

b) one can forget something she believes.

Thus, when we say that a given information is ‘rejected’ we mean that the subject *refused to believe it*, but that this information (and its disbelieving) is perhaps stored in the agent’s mind. When we say that an information has been ‘accepted’ we do not mean that it has just been stored: we mean ‘believed’, taken as reliable on the bases of several factors.

1.2 The Decision to Believe

In our view, there is a basic postulate for our decision[1] to believe:

- *Believe only if you have reasons to believe*

However, a reliable source is by itself a reason for believing. In this way the postulate in not too strong[2], and does not contradict in our view Harman's claim (Harman 1986) and some experimental results (Gilbert 1991). In fact, we don't imply that to believe something one has to logically demonstrate it in her/his knowledge base, or find it strongly ‘plausible’. We just imply that sources give us reasons to believe - quite automatically- and that these reasons have to be stronger than the possible first hand ‘implausibility’ of the new piece of information.

Source Reliability and Belief Credibility

On the one side, believing in a source depends on its assumed reliability[3]. On the other side:

- the Credibility of a piece of knowledge (a candidate Belief) *is a function of its sources*

The basic principles governing Credibility are the following ones :

[1] ‘Decision’ to believe is a strange ‘decision; it is not a true voluntary deliberation, based on computation of advantages and disadvantages. On the contrary (Castelfranchi 1995) it is impossible to consciously decide (and to induce somebody) to believe something on the basis of rewards and advantages or on the basis of threats. We cannot consciously arrive to believe just because it is convenient to believe (Pascal). ‘Decision’ to believe is more a procedural decision.

[2] Otherwise it would be better to assume as a principle: *Believe something if you do not have reasons for rejecting it.*

[3] We cannot prevent ourselves from believing in a source, unless we suspect that there is "something wrong" in it (Castelfranchi 1997). In order to reject the information of a source (even of a social one) we must believe that there is "something wrong" in that source.

  - *1a. If the source is reliable its information is credible and is believed; if it is not reliable its information is not credible and not believed. (In quantitative terms: the more reliable the source the more credible the information provided)*
  - *1b. The many the converging (independent) sources, the more credible the information provided*
- *Any convergent source of knowledge 'confirms' the other* (in particular: S2 confirms S1, when S2 is a new source for a previous item whose source was S1).

Confirmation is a fundamental cognitive 'integration' among sources. It consists of the fact that:

a) after the arrival of a 'confirming source' the item (the belief) is more stable, safe, and certain, and we subjectively are more sure and convinced about it;

b) not only the item is more 'credible', but also the confirmed source is more credible, trusted (it is felt as more 'reliable').

Confirmation is a very important psychological phenomenon: when we control or check something we are just looking for confirmatory sources; proactive behaviour, expectations, and goals imply some 'confirmation' mechanism; there is a very well studied 'confirmatory bias' in our cognition which is not only due to the cost of revising our knowledge (economic motivation) but also to our need for control and to our need to trust our knowledge and our ability to make predictions (Bandura 1982) (apart from self-deception and defence mechanisms).

The 'independence' of the source is clearly very important for our rationality and for resisting social influence. However, from a psychological point of view we have to admit that people are quite ready to accept as confirming and additional evidence also the mere repetition of the same input, and that we do not care so much of controlling the real independence of our social sources (e.g. gossip, tabloids, newspapers). Besides, this might be not so irrational sometimes.

The reliability of the source depends on many different aspects. As for the social or communication sources, many authors identify two dimensions (Fullam 2003; Falcone, Castelfranchi 2004): *competence* and *trustworthiness*. The first is related to the fact that the content of the information is pertaining to a domain the source can be really expert and informed about. For an exhaustive analysis of the second factor, see (Falcone, Castelfranchi 2004).

Other Criteria: Importance and Plausibility

Of course, number and credibility of sources are not the only criteria for accepting/rejecting beliefs. It is not my aim in this paper to examine the specific criteria for belief acceptance and revision, however we need to clarify at least partially which are the other factors that contrast or contribute with the sources to the acceptance/rejection of a belief.

Importance

In our view in the literature on belief revision there is a confusion between two properties of integrated beliefs (Cantwell 1996; Castelfranchi 1996; Paglieri 2004): their *importance* and their *credibility*. They are two distinct dimensions and notions:

- *a belief could be very important but not very credible; or very credible but absolutely marginal.*

By 'important' I mean that it will explain a lot; it will be very *useful for understanding and integrating other information*. 'Credible' means that I have a lot of evidence, sources, supports to believe it. Clearly enough the two aspects are distinct. An integrated belief in a belief network is in fact both supported and supporting: I call *credibility* how much it is supported by external or by internal sources ('plausibility'), and *importance* how much it supports: its explanatory power.

There is a contribution of both 'credibility' and 'importance' in the decision to accept or to maintain a belief. To decide to believe we do not consider only credibility (see later). The fact that a belief is highly 'important' (explains and supports a lot of other beliefs) can be a strong reason for accepting it, and a strong reason for not changing/abandoning it (resistance). In fact abandoning an important belief entails a lot of expensive revisions in our mental map.

Plausibility

To be believed, something should be 'plausible'. Often we resist or reject a new item just because it is not 'plausible'. What is the basis of this kind of evaluation? Why is it a basis for rejection?

Plausibility is *the credibility value assigned* to the incoming item *from inside*. It is evaluated just on the basis of previous knowledge, i.e. the knowledge it has to be integrated with.

Thus, we have two credibility values, one based on external sources, the other based on the niche that has to accept the new item. Metaphorically, one is the value provided by the offering agent that 'gives' the item, the other is the value attributed by the accepting (or refusing) agent. When there is a conflict, and a difficulty to accept the new belief from outside, the conflict is between its 'external credibility' and its 'plausibility'. To accept a new belief the plausibility, i.e. the internal source, should 'confirm' the external source, and converge with external credibility; or, at least, external credibility should be stronger that implausibility.

More precisely, **plausibility** is our attempt to derive, infer the new item from our previous knowledge: the more predicted or expected, the more 'plausible'; in other words we are searching an internal source confirming the external one, in order to better ‘assimilate’ the new information; at least we would like not to have internal reasons for not believing i.e. implausibility. Implausibility is the result of our attempt to derive/infer the opposite of that item. The first attempt is due to our need not only for knowing ‘that’ but for knowing ‘why’ (Aristotle), and for integrating knowledge on such a basis (explanations, reasons). The second attempt is due to *the necessity to verify if there are conflicts* between the new item and the consolidated knowledge. This would be either a reason for rejecting the item or a reason for revising beliefs and for rejecting (dropping) some old items.

In sum, ‘implausible’ means that from my beliefs I might infer the opposite of the incoming data; while there are two level of plausibility: ‘weakly plausible’ means that I do not have conflict, it is not implausible; ‘plausible’ means that I find support in my beliefs in favour of the newcomer, I could also infer it from my previous knowledge.

Of course this is a ‘normative’ perspective; on the psychological side we know that there is experimental evidence that people tend to accept new beliefs without such deep (and long and expensive) controls. Gilbert (1991) for instance has shown that we quite automatically believe all that we comprehend, and that the rejection of ideas comes later as part of a more effortful process. This result should be taken with caution, since Gilbert's notion of ‘comprehend’ is quite broad. Clearly enough we believe quite automatically, by default (Castelfranchi 1995) provided that there are no overt contradictions, immediate implausibility, or previous reasons to suspect/doubt about the source. One should also be careful since in those models the distinction between storing and believing, data and beliefs is not very clear (Castelfranchi 1997; Paglieri 2004).

External ‘credibility’ is just one component of the ‘acceptability’ or ‘believability’ of an item (its properties that will determine whether it will be believed or not). ‘Plausibility’ is another component. But even credibility, plausibility and importance are not enough.

There are also other dimensions that interfere with belief acceptance, like ‘relevance’ and ‘likeability’ (Castelfranchi 1996; Paglieri 2004). By ‘relevance’ we mean how useful and crucial is a given belief for our interests and goals. By ‘likeability’ we mean a special aspect of relevance, i.e. the fact that a belief frustrates or satisfies a goal, is pleasant or unpleasant for us. Our belief adoption is not only a logically rational process; it is also influenced by affective responses - which we cannot address here (see Frijda, Manstead, Bem 2000) - such as wishful thinking, defensive mechanisms and self-deception. Moreover, a given information X could be believable (acceptable for cognitive coherence and withstanding the scrutiny of critical revision), but still remain unacceptable for moral or religious reasons (incompatible with my values and interiorised norms), or very disagreeable and painful to me, given my desires and interests.

2. Belief Support and Goal Dynamics

I will also spend some words on the issue of Intentions adoption and revision, and of its strong relation with belief formation and change.

In general, what Bratman (1990) calls “coherence”, and Cohen and Levesque's (1990) “rational equilibrium” between the agent's intentions and beliefs, is reduced only to the fact that the agent selects and adopts those intentions that he believes to be achievable. 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 instance Rao and Georgeff’s (1991) architecture): some crucial steps, like planning, are not based on beliefs (means-end and causal relations). Only in Bratman, Israel and Pollack’s architecture (1988), beliefs enter all the components of the architecture, determining activation, deliberation, planning, etc. In some sense, I will 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 :

- the ***Processing of goals*****,** from their firing to their satisfaction or abandon: how beliefs determine such a process step by step;
- 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 of changes in a dynamic external environment, or internal cycles of the agent;
- 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.). Also the differences among kinds of goals (like "intentions" vs. "desires", or "expectations" vs. "renounces", etc.) are frequently related to different steps in the goal processing.

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 (revision);

- beliefs determine goal kinds.

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

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

2.1 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 belief of means-end relation which is also the belief of the planning rule (or practical syllogism) and the skeleton 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 either activate or 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; e.g. 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 to achieve 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 do I claim 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 (e.g. an "achievement" goal, (Cohen, Levesque 1991)), specifying the kinds of beliefs one has 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, 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.

Therefore:

- *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);
- *the destiny of an invalided goal 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 the previous assumptions it follows that:

- *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.*

2.2 The Role of Beliefs in Goal Processing

In this section I will assume a very simplified model of Goal Processing - inspired both by psychological models (Heckhausen, Kuhl 1985) and by BDI architectures (Rao, Georgeff 1991) - and I will 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***.

Here I suppose 7 goal states (*Sleeping, Active, Expected, Candidate, Waiting, Chosen, Pursued*), 3 major operators (*Activation, Choice, Planning*), plus several yes/no tests, aimed to check whether or not a given supporting belief is currently associated with the processed goal: failure at different stages results either in demoting the goal to a previous state, or in dropping it altogether, as sketched in Figure 1. Belief support is crucial throughout the whole goal processing, and the aforementioned supporting beliefs (cf. 2.1) can easily be mapped at different stages of the flow, acting as test conditions for passing from a goal state to another.

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 (Conte, Castelfranchi 1995). 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 (Basso, Mondada, Castelfranchi 1993; Castelfranchi 1997); emotional activation of "impulsive goals"; physiological activation of goals (e.g. hunger); goal-Adoption from external requests, norms, commands, etc. (Conte, Castelfranchi 1995); 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).

By way of example, consider the case in which I have a couple of *active goals* (goals I am 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", since 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.

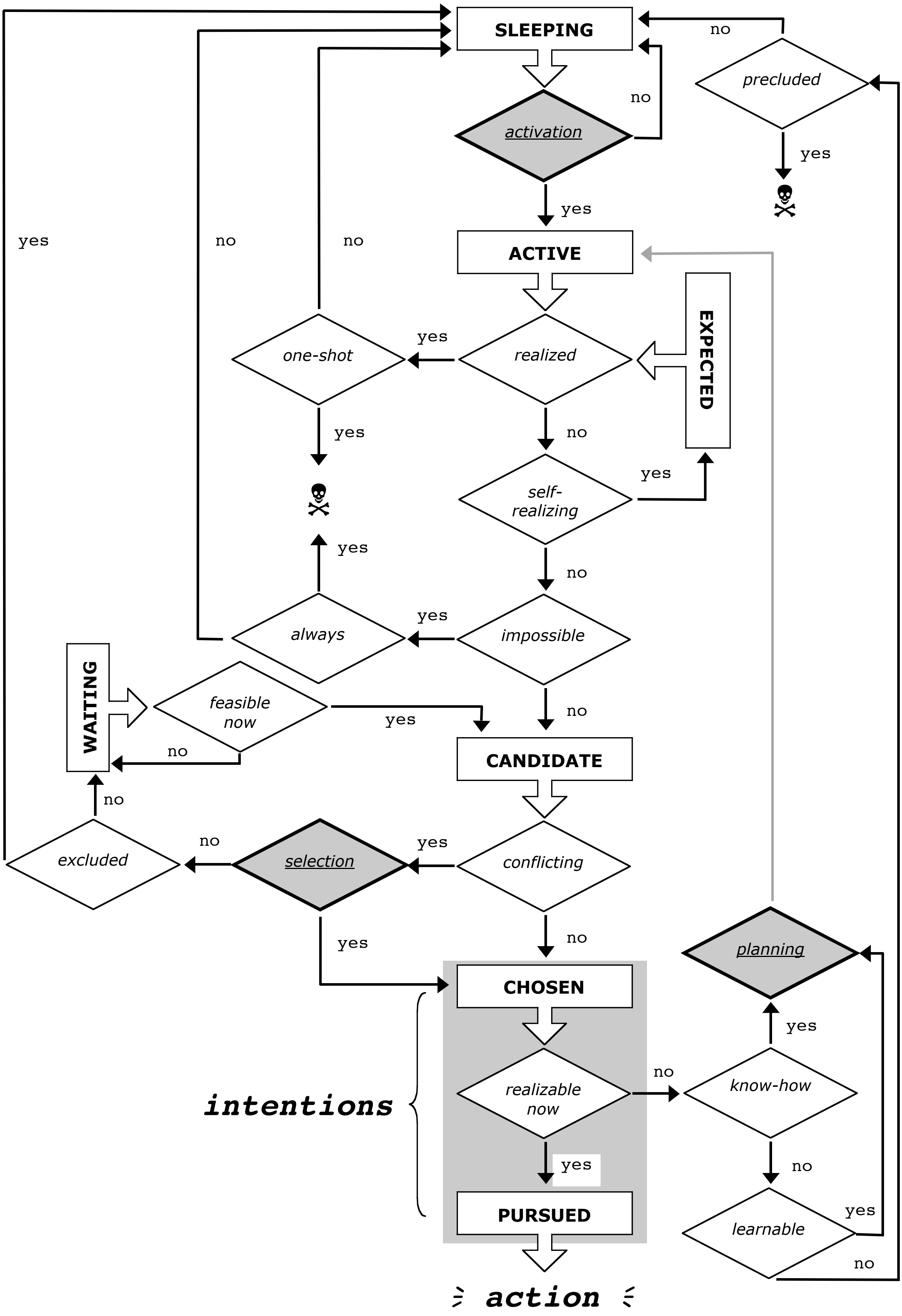

Figure 1. A cognitive model of Goal Processing

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 by adding other "becomes" an intention; and *hierarchically*, from the class of intention which has its typical structure.

2.3 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 and founds 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).

3. Two (In)Concluding Speculations

Reason-Based Believing and Intending as Cognitive Agent's Autonomy

This structure of beliefs-integration and of intention support is responsible of our social autonomy. If one could induce in us any goal or intention at will, we would not have any autonomy. We adopt a given goal from outside and we decide to follow a given intention only on the basis of our own motives and beliefs, and of the reasons they provide us for preferring and planning. In order to influence us to intend to do something, one usually has to modify our beliefs (Castelfranchi 1995).

If one could make us believe everything, he could induce/persuade us to do everything. Fortunately, it is not so simple to make us believe something, just because we believe or not on the basis of evidences and reasons, and usually we have our own multiple and independent information sources. Only if one would succeed in being our only source of information, then he could induce us to believe and to do everything – which is exactly the reason why monopoly of information sources and channels is (or rather should be) such a big social concern in democratic nations.

The Mind in Search for Coherence

Minds search for coherence, both and in parallel at the Epistemic and at the Motivational representations layer (Paglieri, Castelfranchi 2004).

In BDI models of agency (Rao, Georgeff 1991) it has been correctly postulated that an important difference between the level of mere Desires (or wishes) and the level of Intentions, i.e. goals actually directing the agent's behavior, is that while Desires can be subjectively contradictory and the subject can entertain them as such before being obliged to choose, on the contrary Intentions – i.e. what one has decided to pursue and to do – must be subjectively non contradictory. In other words, conflicts between possible/candidate goals must have been solved at the deliberation stage. In our model this process is even more clear, since we assume that Desire or Wishes and Intentions are just *Telos* (Goals, i.e. motivational representation) in different stages of goal-processing: Intentions are just Wishes or Desires (or needs) that we evaluated, preferred to others, sketched for a plan, and committed ourselves to pursue (Bratman 1990). Thus, in our view, we have the same basic representation that start as Desire, then is candidate as possible Intention, then selected as the winner of a conflict (decision making), and eventually adopted as an Intention (with additional features, like the action to be performed for[4]). Exactly the same happens between *data* (information not yet accepted as reliable) and *beliefs* (mental representation to whom the agent is committed): after the selection process, inconsistency cannot be tolerated anymore and contradictions have to be solved. In fact, while data can be and usually are contradictory, beliefs cannot stand contradictions: we cannot believe (at the same time, on the same situation, in the same world, explicitly, and with an high degree of certainty) that p and not-p. So we pass from contradictory data to beliefs by solving such conflict, with some sort of decision making about which data to believe (i.e. which data to promote as belief), either p or not-p or neither of them

This parallel between the processing of epistemic and motivational representations yields a convincing picture of human mind as a *coherence-seeking device* (Figure 2).

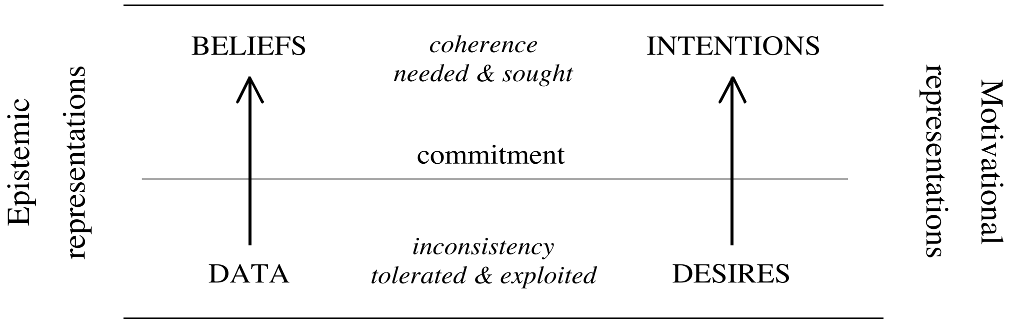
Epistemic representations
BELIEFS
coherence needed & sought
INTENTIONS
Motivational representations
commitment
DATA
inconsistency tolerated & exploited
DESIRES

Figure 2. The mind as a coherence-seeking device

[4] This is the distinction between 'intention that' (about the expected result) and 'intention of' (the planned action for); any 'intention of' presupposes some 'intention that' about some outcome of the action; vice versa, any 'intention that' implies some 'intention of' doing eventually some action, of playing an agentive role.

This parallel also raises a crucial (and still open) question: since goals and intentions are supported and justified by beliefs (and preferred on the basis of beliefs), *what is the relationship between the needed coherence in beliefs and the needed coherence in intentions?*

References

Bandura, A. 1982. “Self-efficacy mechanism in human agency”. *American Psychologist* **37**, pp. 122-147.

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.

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.

Bratman, M.E., Israel, D.J., Pollack, M.E. 1988. “Plans and resource-bounded practical reasoning”. *Computational Intelligence* **4**, pp. 349-355.

Cantwell, J. 1996. *Resolving Conflicting Information.* Technical Report, Department of Philosophy, Uppsala University, June 20, 1996.

Castelfranchi, C. 1995. “Guarantees for Autonomy in Cognitive Agent Architecture”. In *Proceedings ECAI 1994 Workshop on Agent Theories, Architectures, and Languages*, pp. 56-70. Berlin: Springer.

Castelfranchi, C. 1996. “Reasons: Belief Support and Goal Dynamics”. *Mathware & Soft Computing* **3**, pp. 233-247.

Castelfranchi, C. 1997. “Representation and Integration of Multiple Knowledge Sources: Issues and Questions”. In *Human & Machine Perception: Information Fusion*, Cantoni, Di Gesù, Setti, Tegolo (eds.), Plenum Press.

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

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

Doyle, J. 1992. “Reason Maintenance and Belief Revision: Foundations vs. Coherence Theories”. In *Belief Revision*, P. Gärdenfors (ed.), pp. 29-51. Cambridge: Cambridge University Press.

Falcone, R., Castelfranchi, C. 2004. “Trust Dynamics: How Trust Is Influenced by Direct Experiences and by Trust Itself”. In *Proceedings of AAMAS 2004*, N.R. Jennings, C. Sierra, L. Sonenberg, M. Tambe (eds.), pp. 740-747. New York: ACM Press.

Frijda, N. H., Manstead, A., Bem, S. (eds.) 2000. *Emotions and Beliefs: How Feelings Influence Thoughts*. Cambridge: Cambridge University Press.

Fullam, K. 2003. *An Expressive Belief Revision Framework Based on Information Valuation*. MS thesis, University of Texas at Austin.

Gärdenfors, P. 1988. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. Cambridge, Mass.: MIT Press.

Gilbert, D.T. 1991. “How Mental Systems Believe”. *American Psychologist* **46**, pp. 107-119.

Harman, G. 1986. *Changes in View: Principles of Reasoning*. Cambridge, Mass.: MIT Press.

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

Paglieri, F. 2004. “Data-oriented Belief Revision: Towards a Unified Theory of Epistemic Processing”. In E. Onaindia, S. Staab (eds.), *Proceeding of STAIRS 2004.* Amsterdam: IOS Press.

Paglieri, F., Castelfranchi, C. 2004. “Revising Beliefs Through Arguments: Bridging the Gap between Argumentation and Belief Revision in MAS”. In *Proceedings of ArgMAS 2004*, I. Rahwan, P. Moraitis, C. Reed (eds.). Berlin: Springer.

Rao, A.S., Georgeff, M. 1991. “Modelling Rational Agents within a BDI-architecture”. In *Principles of Knowledge Representation and Reasoning: Proceedings of KR91*, J. Allen, R. Fikes, E. Sandewall (eds.), pp. 463-484. San Mateo, California: Morgan Kaufmann Publishers.