

CRISTIANO CASTELFRANCHI, FABIO PAGLIERI

THE ROLE OF BELIEFS IN GOAL DYNAMICS: PROLEGOMENA TO A CONSTRUCTIVE THEORY OF INTENTIONS

ABSTRACT. In this article we strive to provide a detailed and principled analysis of the role of beliefs in goal processing – that is, the cognitive transition that leads from a mere desire to a proper intention. The resulting model of *belief-based goal processing* has also relevant consequences for the analysis of intentions, and constitutes the necessary core of a *constructive theory of intentions*, i.e. a framework that not only analyzes what an intention is, but also explains how it becomes what it is. We discuss similarities and differences between our approach and other standard accounts of intention, in particular Bratman’s planning theory. The aim here is to question and refine the conceptual foundations of many theories of intentional action: as a consequence, although our analysis is not formal in itself, it is ultimately meant to have deep consequences for formal models of intentional agency.

1. BELIEFS, GOALS, ACTIONS: AN ETERNAL GOLDEN BRAID?

It is almost unanimously agreed that beliefs deeply affect our actions, by providing «the maps by which we steer» (Ramsey, 1931). But *why* is it so? A simple answer would emphasize that we have to act in the real world, so that beliefs, insofar as they are truth-aiming mental representations, are necessarily relevant in shaping our actions – they are tools for coordinating our behaviour with the factual features of the environment. However, this simple answer turns out to be simplistic, since it fails to provide a criterion to discriminate between *relevant* and *irrelevant* beliefs in the cognitive regulation of action. Not every belief actually influences our conduct: while arranging one’s wedding in London, beliefs on weather conditions in China do not shape the agent’s actions; conversely, when one is engaged in planning a honeymoon in China, the colour preferences of the bride-to-be about her wedding gown are rather unessential. Whether and how a given belief contributes to determine our behaviour strictly *depends on the goals* we are currently assessing, considering, choosing, or pursuing. To push forward Ramsey’s metaphor, beliefs, like maps, do not influence behaviour in and by themselves: they shape our actions only if, when, and because we decide where to go.

The lesson to be drawn from these rather self-evident considerations is that the influence of beliefs on our behaviour in fact *coincides* with their influence on our intentional processes (or, more

exactly, on motivational processes in general¹). In view of this fact, it is somehow surprising that the connection between belief dynamics and goal processing still remains so poorly understood. In what follows we argue that there is indeed a close relationship between these two types of cognitive dynamics, and we set out to explore how the latter is affected by the former. The main point is that goal processing and intention revision are largely *determined* by belief revision – that is, the process by which we come to intend something crucially *depends on* our assumptions concerning how we consider the world to be, i.e. our beliefs.² More precisely, our aim here will be to detail a model in which specific *types of beliefs* act as *filters* and *triggers* from goal activation to goal execution. According to this view, in order to activate, promote, drop, or suspend a goal, an intention, or an intentional action, one has to provide or modify the appropriate beliefs.

We intend to explore some relevant aspects of this *structural interdependency of beliefs and goals* in cognitive agents, and *its impact on belief change and goal dynamics*. We begin by showing how and why the analyses of these two different kinds of cognitive dynamics in Artificial Intelligence, logic, and (to some extent) philosophy have so far proceeded mostly in parallel, with little mutual integration. In contrast with this practice, we argue that in natural cognitive agents (e.g. humans) doxastic and motivational dynamics are systematically and necessarily integrated. After few preliminary remarks on the conceptual status of the notion of ‘goals’ as used here, we outline its structural correlation with beliefs. In particular, we focus on *the role of supporting beliefs in goal dynamics*, showing that they are necessary to regulate goal processing, determine different goal types, and initiate processes of intention revision. We describe both a taxonomy and a dynamic model of belief-based goal processing, and discuss its impact for a reformulation of a theory of intention – including critical comparison with some standard analyses of intending, e.g. Bratman’s planning theory of intention. Finally, we sum up our results and outline promising topics for future work in the same general direction.

2. GOALS VS. OTHER CLASSES OF MOTIVATIONAL STATES

The notion of belief employed in this essay was precisely characterized and discussed in previous work (Castelfranchi 1997; Paglieri, 2004): here we indicate with the word ‘belief’ any mental representation that is *used* as a plausible substitute for a certain aspect of reality, and that is supposed to be referentially true, i.e. to provide a description that is assumed to correspond, and used as corresponding, to how things actually are. Properly speaking, the notion of belief defines a specific cognitive function that is absolved by a mental representation – in the same way in which a goal indicates a (different) functional role played by a mental representation (see below). In this sense, our theory can be adequately labelled as a form of *cognitive functionalism*: the significant

and heuristic categories of mental representations (e.g. beliefs and goals) are defined by their functions, rather than by their format, content, or other intrinsic features.

Whereas this understanding of belief is not too controversial in the literature (for a survey and critical discussion, see Paglieri, 2006), our notion of goal requires instead to be properly introduced and specified. As it is used here, the class of goals is construed rather broadly, and it encompasses several other concepts familiar both in Artificial Intelligence (e.g. desires and intentions), in cognitive psychology (e.g. motivations, wishes, aims), and in social sciences (e.g. preferences).

Essentially, a goal is defined as an *anticipatory internal representation* of a state of the world that has the *potential for* and the function of (eventually) *constraining/governing the behaviour of an agent towards its realization*. The defining function of goals is to shape, to *direct* in a teleological sense the actual behaviour of the system. When we say that Jack dropped the hair-drier in the bath-tube because he (consciously) wanted to murder his wife, or that Daniel forgot to buy his fiancé a present on her birthday because he (unconsciously) wanted to get rid of her, we are claiming that their actions were *directed towards the realization of some specific state of the world*, and that the anticipatory representation of such state was their goal and the driving force behind their actions.

As for the typical anticipatory nature of goals, this is meant to capture the fact that *goals prefigure a certain state of affairs*, as opposed to representing a state of the world which is believed to be the case. This remains true (and rational) even when a certain goal is already realized: this is the case with the so-called ‘maintenance goals’, i.e. goals about keeping things as they are, which have been sometimes opposed to ‘achievement goals’, i.e. goals about state of things not yet realized (Bratman, 1987; Cohen, Levesque, 1990). This distinction, although significant for some specific purposes, should not hide the fact that both kinds of goal share the same basic anticipatory nature. Imagine I have the goal of being loved by my wife, and let us say that she is in fact deeply in love with me: notwithstanding this happy coincidence, I still maintain my goal – so much so, that I could undertake some actions, like coming home from work at a reasonable hour and doing more often my share of domestic duties, to the purpose of keeping things between us as they are. But, as far as conduct regulation is concerned, this is *not* significantly different from having the goal of being loved by someone that at present does not love me. In both cases my actions are driven and controlled by an anticipatory representation on how the world should be at some time in the future: thus, how the world happens to be now is certainly relevant to choose the most adequate course of (future) action, but it does not change at all the anticipatory mechanism embedded in my behaviour. More generally, the function of a goal is to serve as a frame of reference for driving the agent’s

conduct towards an anticipated state of things. In this sense, goals, either for maintaining the current status quo or for achieving a new one, are anticipatory representations by definition.

Our understanding of goals is partially inspired to the original procedural characterization of ‘goal’ and ‘purposive behaviour’ provided by cybernetics (Miller et al., 1960; Rosenblueth et al., 1968). In this model, a representation of the target state is and remains a ‘goal’ (the original term used in this approach) even when the system is inactive, i.e. it is not working for realizing such a state (homeostatic restoring). These authors assigned to goals several different functions. For instance, the end-state serves as a goal also when the system is just comparing it with the perceptual input, i.e. *evaluating* the world, and when there is not even a mismatch with the world, e.g. because another agent is producing the target world-state.

It is this cybernetic notion of goal that we endorse and refine here: from this standpoint, a goal is a goal even when or if it has not to be actively pursued. Goals (like BDI desires) can be already realized, or they can be self-realizing from the point of view of the agent: they may not require actions by the agent, e.g. due to natural processes or other agents that guarantee their achievement. In this technical sense, it is possible to have the goal that the sun will rise tomorrow, or that Maria will come to visit me on Sunday. Although self-realizing, these states remain proper goals of the agent – and indeed, if the expectations on self-realization result to be wrong, the agent can well decide to directly take action for realizing those goals. For instance, I can decide to go to Maria’s on Sunday, after being informed that she is ill and will not be able to visit me. And for the same token, a goal remains a goal even when the agent lacks the power or the resources to pursue it.

To summarize, a goal is not a representation currently and necessarily orienting and guiding an action; instead, it is a representation endowed with this *potential* function, so that it is somehow ‘destined’ to play this role – but whether or not this role is actually fulfilled depends, as we shall see, on the agent’s beliefs. So construed, goals clearly share many of the properties that BDI models attribute to desires. However, in our view the notion of ‘desires’ is plagued by at least three undesired properties, that should have discouraged its use in this context.

- Desires, dating back at least to Aristotle, have an unavoidable connotation of expected ‘pleasure’: we cannot desire something if it fails to give us some kind of pleasure, or if it entails a prevalent displeasure (although, of course, different people may consider pleasurable different things: consider for instance the desire to die of terminal patients experiencing extreme pain). Hence the use of desire would give to intentional action a hedonistic foundation that is highly disputable, because neither pleasure nor lack of displeasure are necessary conditions for purposive behaviour. This is manifest with respect to artificial agents, that can be goal-oriented without being either pleasure-directed or even

pleasure-sensitive. Besides, the idea that every intentional action is connected with pleasure is arguably false also for humans: for instance, it is at odds with an intentional behaviour determined by norm compliance – unless one is not committed to the controversial proposition (strenuously refused by Kant, among others) that obeying rules necessarily implies a form of moral pleasure akin to hedonistic satisfaction. More generally, not all goals when achieved seem to entail a real felt experience of pleasure: only some goals when realized give pleasure, and thus, while the subject is anticipating such a pleasure, they can be properly ‘desired’ (Castelfranchi, 1998a).

- Desires have an unavoidable mark of ‘endogenous’ pro-attitudes, possibly as a consequence of their hedonistic character: they cannot be induced by external pressure or impositions. On the contrary, a certain goal – for which, from which and around which an intention can be built (see 8) – may originate from a duty or an order, possibly even one that we dislike (Castelfranchi, 1998b). Therefore, not all intentions result from the decision to realize some endogenous ‘desire’ of the agent.
- Desires have a strong connotation as ‘non pursued’ or even ‘non pursuable’ state of affairs. This is one of the reasons in BDI for not using the term in more advanced stages of practical reasoning and deliberation, introducing a new theoretical entity, i.e. intention, as a new primitive. On the contrary, the technical notion of goal used here can be applied both before, without, and after deciding to act, as well as during the persecution of one’s aim and the performance of an action. Thus the notion of goal emphasizes the *continuity* between (what in BDI are called) desires and intentions, opening the way for an operational model of intention formation (see 6 on this important point). Behind the process that leads from a desire to an intention, we claim that there is always a goal from the beginning to the end, which is transformed in its functional properties by subsequent accretions of relevant belief patterns.

Ultimately, the main reason to prefer the notion of goal over that of desire is that it yields a richer model of motivational dynamics. Indeed, since the notion of goal is deliberately designed to cover a wide variety of motivational states, several internal distinctions will be introduced later (see the *typology of goals* discussed in 6), and they will be matched with *different stages in goal-processing*, i.e. the process that brings the agent from a general interest for a potentially relevant outcome to the subjective commitment to bring it about through adequate planning and action. In BDI models (Bratman et al., 1988; Rao, Georgeff, 1991; Wooldridge, 2000; van der Hoek, Wooldridge, 2003), this process is known as *intention formation*, and only two stages are considered: desires, i.e. states of the world that agents find desirable but that they are not yet committed to act upon, and

intentions, i.e. desires that agents are committed to make real through their plans and actions. One of the basic assumptions of our analysis here is that this view of intention formation, although basically correct in its outline, is still oversimplified, and that both formal models and computational systems might benefit from a more refined model of goal processing (see also Dignum, Conte, 1997 for a similar plea, supported by partially different reasons).

3. EXISTING APPROACHES TO THE ROLE OF BELIEFS IN GOAL DYNAMICS

The interplay between beliefs and goals is a core problem in action theory, epistemic logic, and BDI architectures: in fact, some basic elements of the theory of goal dynamics had already been identified, and there are some influential proposals for a formal treatment of these elements (as a case in point, see van der Hoek et al., this issue). However, in our view this area suffers from a disproportion between huge efforts at formalization and implementation, while comparatively small energies have been devoted to cognitive analysis and conceptual clarification. The risk here is that the instrumental apparatus may compromise and hamper the preliminary understanding of some crucial notions. The alternative proposed here is to first clarify a well-defined *ontology* of goals, intentions and commitments, and their processing and structural relationships with beliefs, and only afterwards turn to the problem of formalizing these complex interactions.

To the best of our knowledge, only the minimal core of the theory of belief-goal relations has been so far discussed (Bratman et al., 1988; Cohen, Levesque, 1990; Rao, Georgeff, 1991; Bell, 1995; van der Hoek et al., this issue), while the idea of a *belief structure* supporting goals has frequently surfaced in the study of goal adoption and persuasion strategies (Sycara, 1991; Walton, 1998; Poggi, 2005; Paglieri, Castelfranchi, 2006). But a fine-grained analysis of this structure is still lacking, as well as a precise ontology of different types of belief support and of the multiple roles they play in goal processing.

In fact, what Bratman (1987: 41-42) calls ‘consistency’, and Cohen and Levesque (1990) label ‘rational equilibrium’ between the agent’s intentions and beliefs, are reduced by those authors to the mere fact that the agent selects and adopts those intentions that are believed to be achievable and not in contrast with other intentional commitments. As a result, current BDI models consider beliefs crucial for the adoption or the abandoning of intentions, but their role appears limited and oversimplified in most of these models. For instance, in Georgeff and Rao’s architecture (1991), the belief component cannot be consulted at each step of goal-processing (in contrast to what we will suggest here), and some crucial steps, like planning, are not based on beliefs (means-end and causal relations).³ Only in Bratman, Israel and Pollack’s model of planning (1988) beliefs enter all the components of the architecture, determining activation, deliberation, planning, and so on. With

respect to their work, here we will (1) make explicit the function of beliefs in goal processing, (2) add the idea of their essential ‘supporting’ role, (3) analyze their effect on the typology of a given goal, which depends on its stage of processing, and (4) clarify that the notions of ‘desires’ and ‘intentions’ do not indicate two separate cognitive primitives (contra BDI models and Bratman’s planning theory of intention; see 9), but rather different phases in the processing of goals.

4. PRINCIPLES OF COGNITIVE INTEGRATION: SUPPORT RELATIONS

This analysis of the interaction between belief dynamics and goal dynamics is based on the following principle (Castelfranchi, 1996):

Postulate of Cognitive Regulation of Action:

The goals of a cognitive agent have to be supported and justified by the agent’s beliefs (i.e. reasons). Cognitive agents can not activate, maintain, decide about, prefer, plan for, or pursue any goal which is not grounded (implicitly or explicitly) on pertinent beliefs.

‘To be supported’ here means that it is not just a matter of an already accomplished process, which has generated a given assumption or belief, or a given deliberation in decision-making, or the formulation of a given piece of plan in means-end reasoning. The supporting relation is both diachronic (it happens in time) and synchronic (it leaves traces, a memory of the cognitive path that conduced to the outcome). Thus, the resulting goal remains nested in a given structure of ‘reasons’ (e.g. beliefs and other goals) which has conduced to it and continues to justify it. In such a way these supporting beliefs, by being either tested or just taken for granted, determine the life, the future, and the ultimate destiny of the supported goal. In fact, if a belief which had a role in the preliminary steps of building an intention is later invalidated, the process is stopped and there is either an abortion or a backtracking.

A goal is sustained, both in its current status and in the continuation of its processing, by a rich structure of beliefs, and these beliefs correspond to and keep track of the critical tests that the goal has already successfully passed. This view has two important corollaries:

Corollary 1 (specificity):

At each stage of their processing, goals are filtered or supported by specific beliefs, that determine the properties acquired by the goal in the next stage (e.g. from desires to intentions).

Corollary 2 (dependency):

The destiny of a goal, after that its processing has been compromised, strictly depends on the reasons that caused such a failure (i.e. the specific supporting beliefs that were invalidated).

These principles provide (part of) an operational understanding of one of the basic properties that characterize rational agents, i.e. the capacity of acting *towards* the achievement of their intentions *on the ground* of their beliefs. Here is a good definition and a classical example of this property:

An agent is said to be rational if it chooses to perform actions that are in its own best interests, given the beliefs it has about the world. For example, if I have a goal of staying dry, and I believe it is raining, then it is rational of me to take an umbrella when I leave my house. (...) You would still be inclined to refer to my behaviour as rational even if I was mistaken in my belief that it was raining: the point is that I made a decision that, if my beliefs were correct, would have achieved one of my goals (Wooldridge, 2000: 1).

Other forms of beliefs-goals coordination are of course contemplated, and they imply additional constraints over the agent's beliefs, given its goals (and vice versa): for instance, it is customary to accept that an agent cannot rationally intend p if it does believe p to be already the case, or if it believes p to be impossible (Singh, Ascher, 1993; Wooldridge, 2000: 9, 25). Later on we will show that these standard classes of supporting beliefs are far from being exhaustive, and that a more detailed typology of the different beliefs necessary to support rational goal processing would be indeed valuable. However, first we need to define more precisely the basic tool applied in this analysis: *support relations*.

We talk of 'support' relations whenever a cognitive item (belief or goal) holds *because* of such relations, and only *as long as* they stand in place. In the case of a goal, there is a specific structure of beliefs that is necessary to maintain and justify it (*goal-supporting beliefs*): in turn, these beliefs have their own relative structure of justifications and supports (*belief-supporting beliefs*), that was discussed in detail elsewhere (Castelfranchi, 1997; Paglieri, 2004; 2006). To define a set of beliefs as 'supporting' a given goal means that, without such beliefs, the related goal would be dropped from its current state, and would change its nature and functional properties.

Imagine for instance you went to the airport with the intention of taking a certain flight to reach Oslo, but then you discovered that such a flight was not heading there. Here your process of belief revision is the prime responsible of your rational decision to drop the intention of taking that flight, since your previous intention was based on the (mistaken) belief that doing so was instrumental to reach your final destination. On a slightly different and more cheerful note, suppose that in the same initial situation you discover instead that there is another flight to Oslo arriving at destination earlier than the one you had chosen before, and you can switch flight without losing any money. Again such a change in your beliefs would make it rational for you to abandon your previous intention, as long as you believe that the new available option is to be preferred as a means to achieve your ultimate goal.

These general observations just serve to frame a fundamental question, which is addressed in the next section: *Which ones*, among the many beliefs relative to p (or to goal p), are actually

supporting it? In other words, is it possible to provide a well-defined characterization of goal-supporting beliefs within the broader category of goal-related beliefs?

5. GOAL-SUPPORTING BELIEFS: A TENTATIVE ONTOLOGY

In this analysis we are interested to study those beliefs that support the processing of a given goal, starting from its activation and ending up with its instantiation in a present-directed intention. What we call *goal processing* can be seen as a description of the process that leads from desires to intentions, in BDI jargon – and in fact we will argue that one significant benefit of this kind of analysis is to fill in this major explanatory gap of BDI, describing *under what (doxastic) conditions a desire becomes an intention*.

The following is an attempt at drawing a comprehensive list of different *types* of goal-supporting beliefs, arranged according to the different *functional role* they play in goal-processing, from activation to action:

- *Motivating beliefs*: aside from the case of bodily activation of goals (e.g., a feeling of hunger which triggers my search for food) or emotional arousal (e.g., a sudden sensation of fear that generates the goal of running away), goals are often activated by beliefs on the current state of the world. Here two sub-classes are considered:⁴
 - *Triggering beliefs*: beliefs that reactively activate goals on the basis of a pre-established association. Example: it is my belief that the fire alarm is ringing that activates my goal to escape from my office.
 - *Conditional beliefs*: beliefs that activate a goal on the basis of the conditional nature of the goal itself. Example: it is the belief that today is Sunday that activates my conditional goal of going to mass on Sunday.
- *Assessment beliefs*: in order to consider a goal as candidate for being pursued, I cannot believe that such a goal is either already realized, self-realizing, or plainly impossible. Here we propose to distinguish between these different sub-cases:
 - *Self-realization beliefs*: beliefs concerning the fact that one of my goals will come to be realized in the world autonomously and without my direct intervention. Example: it is the belief that my salary will be paid in due time by the University that prevents me from taking direct action to ensure that my goal of being paid is realized, unless I have reason to suspect that the University will not pay (i.e., unless the self-realization belief is dropped or weakened).

- *Satisfaction beliefs*: beliefs concerning the fact that one of my goals is already realized, and that it will remain as such without my intervention. Example: the belief that I am married to the woman I love prevents me from pursuing the goal of marrying her.
- *Impossibility beliefs*: beliefs concerning the fact that one of my goals is impossible at a given time, or it will never be possible. Example: the belief that an obscure and rather unknown political party will never win the election can effectively prevent voters from supporting that party, notwithstanding their desire to do so.

It is worth noticing that the absence of these beliefs implies that, in order to bring about the goal, the agent both *can* and *has to* do something. The goal is within the agent's reach, but it is also up to the agent (if he decides that way) to do something for achieving it – it has not yet happened, and it is not expected to happen unless the agent does not contribute to make it occur. Only after this crucial stage of belief-filtering goals become proper candidates for being chosen and intended by the agent: i.e., they become *pursuable*, not only in the sense of 'possibly feasible', but also implying that such goals, if chosen over other alternative aims (see below), will require the active engagement of the agent.

- *Cost beliefs*: beliefs concerning the costs that the agent expects to sustain as a consequence of pursuing a certain goal, in terms of the necessary resources that will be allocated to that end.⁵ Example: my belief concerning the time I will need to prepare an article for a famous journal is relevant to make me decide whether or not I should pursue such a goal.
- *Incompatibility beliefs*: beliefs concerning different forms of incompatibility between different goals, that can force the agent to choose among them, either in absolute terms or for the time being.⁶ Given two goals G_1 and G_2 , I can believe them to be incompatible in several different ways: either because the results of G_1 and G_2 cannot be both true in the same world (*conflicting aims, or terminal incompatibility*), or because I cannot achieve both G_1 and G_2 at once (*conflicting resources, or instrumental incompatibility*), or because both goals are mere means to the same end (*superfluity, or convergent means*). Example of terminal incompatibility: it is my belief that I cannot be a Catholic priest and married at the same time that prevents me from trying to pursue both these goals, even if I desired to do so. Example of instrumental incompatibility: it is my belief that I do not have enough money to buy both a new car and a new bedroom that force me to choose what to do first, although I desire both of them. Example of superfluity: my belief that paying in US dollars and paying in EC dollars are both sufficient means to buy the fancy Caribbean T-shirt that I cherish so much is what

prevents me from doing both things at once while shopping in Antigua, although both are valid sub-goals to my general plan.

- *Preference beliefs*: beliefs concerning what (incompatible) goals should be given precedence over others in the current context. These beliefs on goals are in turn grounded on the integration of at least two sub-classes:
 - *Value beliefs*, concerning the subjective value of a certain goal, given my current interests. Example: my belief that completing my long overdue PhD dissertation is more valuable for me than submitting an article to a famous journal can make me decide that the former goal is to be preferred to the latter.
 - *Urgency beliefs*, concerning when (if ever) a given goal will ‘expire’, i.e. it will be no more possible to achieve it. Example: my belief that the deadline for submitting an article for a prestigious special issue is closer than the deadline for completing my PhD dissertation can persuade me to prefer the former goal over the latter.
- *Precondition beliefs*: beliefs concerning the necessary preconditions for successfully pursuing a given goal by executing the appropriate action. Here we distinguish two sub-cases:
 - *Incompetence beliefs*: beliefs of ‘internal attribution’ (Weiner, 1974), self-efficacy, and confidence; they mainly concern both the basic know-how and competence, and the sufficient skills and abilities needed to reach the goal, given my convictions on how the goal can be reached. Example: I cannot start to cook the liver Venetian style if I believe I am ignorant of this fashionable recipe, or I believe to be unable to successfully putting it in practice.
 - *Lack of conditions beliefs*: beliefs of ‘external attribution’, concerning external conditions, opportunities, and resources; they cover both conditions for the execution of the appropriate actions, and conditions for the success of a correctly performed action. Example: if I believe I have no onions, or I believe that my stove is out of order, I cannot start cooking the liver Venetian style, no matter how much I would like to.
- *Means-End beliefs*: beliefs concerning the instrumental relation between a given goal and an action or an event which is considered to serve to achieve the former, and therefore can be assumed as a means (sub-goal) to that end.⁷ Example: it is my belief that one needs a good pan to cook the liver Venetian style that compels me to look for a professional pan, given my goal of cooking a wonderful liver Venetian style for my friends.

As far as different functional types of goal-supporting beliefs are concerned, this list is intended to be as exhaustive as possible. However, it is not meant to be complete with respect to single specific beliefs, since this would clearly be a hopeless task. Our claim is rather that, for a goal to be in a certain stage of processing, it is *necessary* to also have (or lack) some beliefs from each of these categories (more on this in the next section). This conveys the idea that there are several interrelated and yet different stages in goal processing where beliefs play *a necessary role* – in fact, in this model goal-supporting beliefs act as *tests* or *gates*, determining whether or not a certain goal is suitable for the next stage of processing (see 6). While it is possible in principle to distinguish different types of beliefs according to their role in this process of goal selection, it is important to stress that a given goal is always supported by a *structure* of beliefs, rather than a mere list.

It is also worth noticing that there are significant interactions among some of these goal-supporting beliefs. For instance, an impossibility belief may be derived from a belief about an expired deadline. Means-end beliefs, although strictly necessary only to pursue an already chosen goal, can enter the process also at previous stages, e.g. they can serve to assess the effective costs of a given course of action and to compare it with other available options. At a general level, this implies that goal-supporting beliefs may be in structural relations among them: they may mutually support each other, or be components of more complex beliefs.

Finally, the interest of this goal-supporting belief structure is not merely topological, but it concerns instead the *function* that such beliefs serve in goal dynamics. In the next section, we provide an outline of a dynamic model of goal processing, with special emphasis on the crucial role played by goal-supporting beliefs.

6. GOAL-SUPPORTING BELIEFS: A DYNAMIC MODEL

The ‘dynamic twist’ in the ontology of goal-supporting beliefs just discussed is provided by the fact that each of those types (1) *intervenes at a different stage* in goal processing and (2) *has different effects* on goals. Below is a schematic reconstruction of how the whole process is supposed to work, to be further detailed in what follows.

Goal Type	Process Stage	Supporting beliefs	Beliefs sub-classes	+/-
Active Goals (= desires)	ACTIVATION	Motivating beliefs	Triggering beliefs	+
			Conditional beliefs	+
	EVALUATION	Assessment beliefs	Self-realization beliefs	-
			Satisfaction beliefs	-
			Impossibility beliefs	-

Pursuable Goals			
	DELIBERATION	Cost beliefs	-
		Incompatibility beliefs	-
		Preference beliefs	Value beliefs + Urgency beliefs +
Chosen Goals (necessary for future-directed intentions)			
	CHECKING	Precondition beliefs	Incompetence beliefs - Lack of conditions beliefs -
		Means-end beliefs	+
Executive Goals (necessary for present-directed intentions)			
ACTION → Feedback and subsequent (1) belief revision and (2) plan diagnosis			

This scheme operationalizes the transition from ‘BDI desires’ (active goals) to present-directed intentions (in which executive goals play a crucial role) – that is, the passage from a potential fancy to an intentional action. Moreover, it highlights the crucial role played by (different kinds of) beliefs throughout the process. *Motivating beliefs* are possibly responsible for goal activation, i.e. they determine whether or not a potential goal holds an actual interest for the agent in the present circumstances – but notice that also other, non-doxastic patterns of activation may intervene at this stage (e.g. physical arousal, emotional triggering). *Assessment beliefs* scrutinize whether or not an active goal can in principle be pursued, and whether it needs to be pursued if one wants to make it happen (goal evaluation), i.e. whether or not the goal constitutes a well-formed object for deliberation and a candidate to become an intention. *Cost beliefs* determine whether pursuing the goal is, in and by itself (i.e. prior to any conflict with other goals), worth the effort in terms of resources that will need to be allocated. As for *incompatibility beliefs*, they set the scene for the actual choice, assessing whether there are pursuable goals that stand in conflict with each other, so that a resolution (if needed) can be made on the ground of *preference beliefs*, determining what goals are actually chosen by the agent in the current situation. Finally, goal checking depends on *precondition beliefs* and *means-end beliefs*: if the former are satisfied (i.e. the agent neither believes to be incapable of performing the action required to achieve its goal, nor believes that the factual preconditions for success do not hold), then the chosen goal can be directly instantiated into an action; otherwise, means-end beliefs are consulted to check whether it is possible to devise a plan to make the chosen goal achievable in the end.

Some additional remarks are needed, before considering further merits and limits of this proposal. First, some steps in goal processing require as a necessity only the *absence* of a given

belief (*negative filter*), while in other cases the *presence* of a given belief is instead required (*positive filter*). In the scheme outlined above, this is indicated with ‘+’ and ‘-’: assessment beliefs, cost beliefs, incompatibility beliefs, and precondition beliefs are all instances of negative filter, since their absence is sufficient to let the procedure progress;⁸ in contrast, motivating beliefs, preference beliefs, and means-end beliefs constitute positive filters, because here the presence of a given belief is needed to carry on the process. Some of these beliefs will be explicitly tested and verified within the current knowledge base of the agent, whereas others will be just taken for granted and assumed by default. All of them, however, remain as crucial supports for the goal: invalidation of either a negative filter (e.g. I realize that my goal is already achieved) or a positive one (i.e. I discover that I do not have the means to achieve the goal) forces reconsideration of goal processing and impinges on the goal destiny.⁹

Second, although the sequence of the different stages may allow for some degree of flexibility (e.g. anticipating means-end reasoning before or during deliberation), each stage implies that *all* the previous belief tests have been either passed or taken for granted, and that they hold here and now for the agent. This requirement constrains the definition of all goal types, so that the labels we use to name them are just that – labels, technical devices conventionally used as a shorthand for indicating a goal plus the characteristic pattern of supporting beliefs that it acquired going through all the previous stages of processing. In particular, let it be understood that, in what follows, the label ‘chosen goals’ (*C-goals*) is a short-hand for goals that are ‘chosen for pursuit’, i.e. goals that are chosen after having already been assessed as pursuable and up to the agent. We decide to name them ‘chosen’ both for brevity, and to emphasize that, at this stage, they are the outcome of a (particular type of) choice. Similar considerations apply to all other goal types, as what follows will illustrate.

Having outlined the necessary limitations of this scheme of belief-supported goal processing, let us now turn to its merits. On the formal side, it immediately offers some suggestions on how to *reduce the number of motivational primitives* in BDI logics from two (desires and intentions) to one (goals), at the same time providing some insight on their mutual relationship. In fact, once we take active goals (akin to BDI desires) as our primitive of choice, we can then define all other goal types, down to executive goals, in terms of such a primitive plus presence / absence of some specific beliefs, roughly as follows:¹⁰

Active goal (desire): *GOAL (p)*

Pursuable goal: *P-GOAL (p) =*

GOAL (p)

AND *no assessment belief on p*

Chosen goal (necessary for future-directed intention): **C-GOAL (p)** =

P-GOAL (p)

AND *no cost belief on p such as to prevent pursuing it*

AND *no incompatibility belief on p*

OR *no goal r preferred over p given preference beliefs*

Executed goal(necessary for present-directed intention): **E-GOAL (p)** =

C-GOAL (p)

AND *no precondition belief on p*

This theoretical proposal stands in sharp contrast with the dominant view of intentions, i.e. Bratman's planning theory on intention, in several respects. Moreover, further discussion of what exactly is the place of intentions in our theory of belief-based goal processing is needed. Both these issues will be carefully considered, respectively in section 9 and section 8. First, however, we want to introduce an important addition to the model: a theory of *intentions in agenda*, i.e. goals that have been chosen, but for which the conditions for execution do not yet hold.

7. INTENTIONS IN AGENDA

After a goal has been preferred to others and chosen, so that the subject has decided to do some appropriate action to bring it about, the relevant conditions for executing the action and/or for realizing the goal can or cannot be immediately satisfied. If these conditions are present or shortly forthcoming, the intended action is put in execution and the goal becomes currently and actively pursued by the agent. If instead the conditions for a successful execution are not yet there, the goal of doing an action is put into a very special 'waiting room': we call it the agent's *agenda*, in the literal meaning of the word – actions to be done. These chosen goals are shelved for the time being, but are not meant to be re-examined or reconsidered later: they are just to be executed when the expected conditions will be there. Thus, either explicitly or implicitly, a future-directed intention, or intention in agenda, is a chosen goal (on myself in the future) to have a present-directed intention, or intention in action. I have *now* the goal of having *later* the already decided goal of performing the action. The agenda is the memory of such goals, and serves as a reminder and a commitment to myself. The very act of putting a goal in agenda is *per se* the formulation of this binding on the future, and acts as a message and prescription to myself¹¹ (more on the connection between future-directed and present-directed intentions in 8). Like we do with our real agenda, we annotate in our cognitive agenda the things to do in the future, given our current intentions. And just like our real agenda, our cognitive agenda is a relevant tool for inter-temporal coordination of our actions in view of some long-term purpose (Bratman, 1999: 58-90).

More specifically, there are several conditions that a chosen goal must meet, in order to be formulated as a future-directed intention and put into our agenda. With reference to the content of that goal, the agent has to:

- (i) believe that the conditions for the performance of the action or for the achievement of the goal after performance of the action are *not currently true*;
- (ii) believe, or better *expect*, that they will be *true later*, in time for realizing the goal;
- (iii) believe/expect that he will not change his mind, since he has already chosen and *decided* what to do, and a specific time-slot in the agenda has been allocated and committed to the intended action, possibly together with other resources (e.g., attentional focus, prior planning, etc.);
- (iv) as a consequence of (ii) and (iii), the subject believes/expects (possibly with some uncertainty) that he will *do the action* – which is different, with reference to the discussion above, from believing that the action will be necessarily successful.¹²

The role of intentions in agenda is crucial to ensure the rationality of planning agents, along the lines described by Bratman: they affect future choices, resource allocation, coherence in planning, and so on. Before considering how these important functional properties of intentions are accounted for by a belief-based model of goal processing, we must address a crucial question that so far what merely hinted at: *What is an intention*, in the terms of our theory of goal processing? Are intentions mere chosen goals, i.e. goals that have successfully passed all the relevant belief tests up to the execution stage, or is there something more to them? The next section addresses this important point, outlining a theory of intentions as *double-faced teleological attitudes*.

8. GOALS AND THE DOUBLE-FACED NATURE OF INTENTIONS

As it was made clear by the previous analysis of goal processing, in our view a necessary element of an *intention* is a goal in special conditions, after a specific elaboration, supported by a particular frame of beliefs. Conversely, a goal *becomes* an intention only after passing a series of screening tests, in which specific beliefs act as filters. This applies both to *future-directed* and *present-directed* intentions. A future-directed intention requires a *chosen goal*, i.e. a goal that is active in the agent's mind, that it is not believed to be either already realized, self-realizing, or impossible, the costs of which are estimated to be bearable by the agent, and (most crucially) that it has been preferred over other competing alternative ends, but that is not immediately realizable. Instead, a present-directed intention is built from an *executive goal*, i.e. a chosen goal that is immediately realizable, for which all the necessary preconditions hold, no further means is needed, and temporal priority is given over other chosen options (if any). This view entails that what makes the difference

between a mere goal and an intention, and between future-directed and present-directed intentions, is the stage of goal processing to which each of them belongs, which in turn is determined by, and reified in, a particular frame of supporting beliefs. Hence, *intention can be precisely defined and analyzed in terms of goals and beliefs* (in contrast with some tenets of Bratman's planning theory of intention; see 9), although their interaction originates several emerging functional properties that are characteristic of intention and intention only.

However, when we claim that an intention *requires* a goal at a specific stage of processing, this is very different from saying that such a goal, in and by itself, *is* the corresponding intention. Things are more complex than that, because at the very moment when a goal becomes an intention, a crucial transformation occurs in our mind: a chosen goal, i.e. a goal that we have elected among other to be pursued, immediately becomes, in effect of our choice, a double-faced entity, which includes both a *target* (what we wanted to achieve in the first place) and a *vehicle* (the action or plan that will achieve it). We propose to define an intention as the *combination* of these two different teleonomic objects, and we maintain that goals, as well as other motivational states in general (desires, wishes, fancies, interests), do not share this functional structure, which is therefore characteristic of intentions. We can desire to win the Nobel prize without desiring to undertake the necessary efforts to achieve it; but we cannot intend to win that prize, if we do not also intend to do something about it (i.e., whatever we consider to be necessary to achieve this aim).

This view is closely connected with the distinction between *intention-that* and *intention-to*. Here is the classic, insightful formulation of this distinction, as provided by Wilfrid Sellars:

Intentions are not limited to intentions *to do*, whether now, or later, or on the condition that a certain circumstance obtains. There are also intentions *that something be the case*. The latter, however, are *intentions*, practical commitments, only by virtue of their conceptual tie with intentions *to do*. Roughly, "It shall be the case that-p" has the sense, when made explicit, of "(*Ceteris paribus*) I shall do that which is necessary to make it the case that-p" (1967: 1-2).

Sellars argues that an intention-that, once carefully considered, is ultimately dependent upon an intention-to. We want to further elaborate on this point, claiming that every intention-that *entails*, *generates*, and remains fundamentally *connected with* an intention-to that serves as the vehicle through which the intended aim is achieved. More specifically, what we are suggesting here is not to consider intention-that and intention-to as just two different types of intentions, as it is customary in the literature, but rather as the *two necessary elements of any intention*. In other words, we propose to reserve the technical term of 'intention' to the functional combination of an intention-that and an intention-to: an anticipated, motivating result combined with a practical action (or sequence of actions) that realize such a result.

This two-faced, means-end structure of intentions applies to both future-directed and present-directed intentions (with some important differences, that we will discuss in a moment), and it can be spelled out as follows: whenever an agent has the intention of doing something intentionally, this requires both the *intention-to perform that action* (Int-Act) and the *intention-that one of the expected results of the action will hold* after execution (Int-End). Analogously, if an agent has the intention-that a certain result obtains (Int-End), this necessarily requires a corresponding intention-to do something (possibly still unspecified at this stage; see Bratman, 1987: 29-30) in view of that end (Int-Act). Each of these ‘intentions’ can be analyzed in terms of goals at a given stage of processing, with their characteristic frame of supporting beliefs, but *it is only the combination of the two of them that captures the exact meaning of intending*. If we consider first the case of future-directed intentions, we see that what the agent includes in the agenda is actually a compound of:

- a *chosen goal that p* (future-directed Int-End), defined along the lines discussed above;
- a *chosen goal of doing A* (future-directed Int-Act), also characterized in terms of goal processing and supporting beliefs.

Moreover, the agent must be *aware* of the means-end relationship between doing A and achieving p, i.e. he must have:

- *the belief that doing A is a means to bring it about that p*.

This belief on the instrumentality of Int-Act for Int-End is crucial to the notion of intending, as the following example makes clear. Imagine that Adam, a young PhD student, intends to embarrass publicly his colleague Eve, e.g. because he envies her academic success. One day, Adam has to send one of Eve’s papers to his supervisor, and he does so intentionally, after being asked for that particular paper by his supervisor, who became interested to know more about Eve’s work. However, the paper happens to be scientifically very poor (a fact of which Adam was completely unaware), so that the supervisor writes a terrible review on it for a famous philosophical journal, which results in a huge public embarrassment for Eve, to Adam’s joy. But would we say in this case that Adam intentionally embarrassed Eve? Certainly not. And yet, here he had the intention of sending the paper to his supervisor (Int-Act), which turned out to be a means to achieve his intention that Eve should suffer public contempt (Int-End). But the crucial point is that in this case Adam did not conceive his intentional action as instrumental to his final end, i.e. he had no belief concerning the relevant means-end relationship – and this is precisely what makes his (successful) action of embarrassing Eve utterly unintentional.

Notice that these three conditions, i.e. Int-Act, Int-End, plus the belief on Int-Act being a means for Int-End, are not only *necessary* for intentional action, but also *jointly sufficient* – that is, we do not need to postulate anything more. In particular, we do not need to assume that Int-End is

the predominant reason for developing Int-Act, so that the agent has Int-Act *because* of Int-End. This would be too strong. Consider again Adam's case, and imagine that now Adam, while sending Eve's paper to his supervisor, is perfectly aware of its poor quality, and even hopes that the whole thing will result in Eve being put down in public. Here it is clear that the action of embarrassing Eve is performed intentionally by Adam. And yet, it would be misleading to stipulate that Adam's intention of sending the paper to his supervisor is generated, or even predominantly motivated by his intention of embarrassing Eve. In fact, it is not: Adam formulates the intention of sending over the paper for another reason – namely, that he was asked to do so by his supervisor, and he could not have refused, even if he would have wanted to. So it is enough for Adam to believe that this action fits perfectly within his scheme on Eve, for making the whole act of embarrassing Eve intentional, without any need of postulating that he sent the paper because he wanted to embarrass her.

Now it is time to recall that intention-that and intention-to, interpreted as sub-components of a proper intention, are *reducible to goals at a certain stage of processing* – chosen goals for future-directed intentions, executive goals for present-directed intentions. This is crucial, because otherwise our definition of intention would seem to entail infinite regression. On the contrary, we have now enough elements to provide the following operational definitions of, respectively, future directed intentions (FDI) and present-directed intentions (PDI).

FDI (p) iff $C\text{-GOAL}(p) \ \& \ C\text{-GOAL}(A) \ \& \ Bel(A_means_for_p)$

PDI (p) iff $C\text{-GOAL}(p) \ \& \ E\text{-GOAL}(A) \ \& \ Bel(A_means_for_p)$

With reference to future-directed intentions, what happens in intention formation is that, after the deliberation phase, we put in agenda both the C-goal-to-do-something (not necessarily specified) in order to bring about a C-goal-that-something, and we consider ourselves committed to *both* goals when the appropriate circumstances will arrive – we regard them as (defeasible) prescriptions, commands, or imperatives (Castañeda, 1975).

As for present-directed intentions, the basic mechanism is still the same, because intentional action maintains in execution the layered structure of the future-directed intention that triggered the behaviour: it presupposes not only the present-directed intention to perform the action, but also the intention to bring it about at least one of its results – more precisely, the action elicited by the present-directed intention is a *means* to achieve the result which is the object of the intention (the original chosen goal). In other words, there is an intrinsic instrumental element in the performance of intentional acts, that must be accounted for in our analysis. More precisely, in present-direction intention we have an E-goal-to-do something coupled with an *expectation* on the motivating results

of the action, so that the action is for something, in order that something is the case. This is most adequately described as an expectation, rather than a further E-goal or present-directed intention-that, because the very notion of present-directed intention-that is problematic, and in our view self-contradictory. In fact, a goal-that either concerns some future outcome (possibly very close in the future, but anyway not yet present), therefore it is not present-directed; or it is truly in the present, therefore it is already realized, hence it cannot be intended, because I do not have to do anything for making it happen (the so called maintenance goals are no exception here, since they are goal-that something will be maintained, i.e. it will be in the future as it is right now; see 2 for details). Properly speaking, present-directed intentions cannot be intention-that, by the very definition of what an intention is. In contrast, they are intentions-to (in order that), i.e. intentions in action, and the expected future outcome that motivates the action absolves one of the function of teleonomic representation: it is to be matched against the world to check whether the action was successful or not. But the real motivating power behind my intentional behaviour is provided by the corresponding future-directed C-goal-that, from which the present-directed E-goal-to was generated and depends. This is why in the schematic formulation of PDI presented above the E-goal(A) is coupled with a C-goal(p), whereas A is the action considered instrumental to achieve p.

An important footnote to this analysis is that, in order to intend that something is the case, it is *not* necessary to have already a detailed plan on how to bring about this intended result, with all the corresponding intentions to do the relevant actions for that final end. It is sufficient (and necessary) to maintain that such a result depends also on the agent, that there is an action or plan for achieving it which is known or can be worked out, and that the agent will probably be able and in condition to execute it. According to the well-established view that partial plans are filled-in only when the time of actual action approaches (Bratman, 1987; Bratman et al., 1988; Cohen, Levesque, 1990), the real intended actions need to be fully specified, especially for the executive motor actions, only at the very moment of their execution, and on the basis of contextual data.

9. BEYOND BRATMAN: WHY INTENTION IS NOT A PRIMITIVE

In his seminal work *Intention, plans, and practical reason* (1987), Michael Bratman develops a *planning theory of intention*, in which the notion of intention is closely intertwined with the notion of planning, and at the same time it is treated as being irreducible to any compound of beliefs and desires. This view is presented as a radical alternative to the *desire-belief model* (Anscombe, 1957; Goldman, 1970; Audi, 1973; Davidson, 1980; Davis, 1984), according to which intentions can be described in terms of beliefs and desires. In contrast, Bratman aims to establish intentions as «distinctive states of mind, on a par with beliefs and desires» (1987: 20). In his analysis, as well as

in all the BDI formalisms built on it (Georgeff, Lansky, 1987; Bratman et al., 1988; Rao, Georgeff, 1991; Wooldridge, 2000), this results in *treating intentions as a primitive notion*, in parallel with beliefs and desires. Our approach differs sharply on this point, since we take intention to be ‘a distinctive state of mind’ that is *precisely definable in terms of goals and beliefs*, along the lines discussed above.

In particular, what we propose is a *methodological, non-eliminativist reduction*, according to which (i) intentions do exist as specific and relevant mental states, that (ii) happen to be formed by complex structures of simpler, atomic notions, namely goals and beliefs, so that (iii) their characteristic properties can be analyzed as an emergent effect from the interaction of their cognitive components. This last point is especially important: in our view, intentions do have distinctive functional features, that are not shared by any other motivational states, such as desires / active goals. However, these features are produced by the internal structure of intentions in terms of goals and beliefs, and it is only by looking at the way in which these elements interact that we can really explain the properties of intention, instead of merely listing them. To borrow an obvious metaphor, it is quite evident that the molecule of H₂O has distinctive properties of its own, and yet it is equally obvious that such properties are the emergent effect of the interaction of its constitutive atoms – so much so, that a step change in chemistry was made precisely when a connection was worked out between the properties of molecules and their atomic structures. Similarly, we defend the idea that intentions are cognitive molecules, rather than atoms, and that a much richer understanding of intentional behaviour is achieved by looking at intentions in terms of goals and beliefs.¹³ So the point is not whether intentions are distinctive mental states worth investigating (we certainly agree with Bratman that they are), but rather whether intentions should be analyzed as either primitive or derivative notions – and here we strongly disagree with Bratman, since he maintains the former view, while in what follows we champion the latter.

However, by defending the idea that it is not necessary to take intention as a primitive notion, we are *not* claiming that beliefs and goals alone are enough to capture intention as a derivative notion: they are certainly necessary ingredients, but they are unlikely to be sufficient – as our discussion of the special role of the agenda (see 7) suggested, as a possible way to operationalize the concept of *commitment*. Although in-depth discussion has to be left to other works, we are fairly convinced that it is impossible to have a satisfactory theory of intention without adding something else to the structural relationship between beliefs and goals. But this is not a reason to make intention a primitive. More generally, this additional ‘something’ that is still missing does not seem to be yet another kind of mental states, but rather a fundamental property, such as commitment (Cohen, Levesque, 1990).

Notwithstanding our disagreement with Bratman concerning the primitive vs. derivative nature of intentions, we accept by and large his functional analysis of the properties of intentions. This poses a challenge to our reductionist approach, insofar as it must be capable of capturing all the distinctive features of intentions identified by Bratman. Most noticeably, these include:

- the fact that intentions are *conduct-controlling attitudes*, whereas desires (and goals) are merely potential influencers of action;
- the fact that intentions have a certain amount of *inertia*, so that agents show a characteristic resistance to drop or revise them;
- the fact that intentions have the function of *constraining future reasoning*, both by stimulating the agent to find proper means to achieve them, and by preventing the agent from intending other things that are in contrast with current intentions – the so called ‘screen of admissibility’ effect.

A proper model of intention must acknowledge these facts, as Bratman correctly points out. As for intentions being conduct-controlling attitudes, this feature is built in our model, where only executive goals, i.e. constitutive elements of present-directed intentions, directly control the immediate behaviour of the agent, whereas chosen goals, i.e. the building blocks of future-directed intentions, influence action by generating present-directed intentions on the means to achieve these goals. In contrast, goals that are not (yet) chosen do not have the power to control the agent’s behaviour, consistently with Bratman’s observation on the difference between desires and intentions in the cognitive regulation of action.

Similar considerations apply also to the constraints that having an intention poses for future reasoning – that is, the third property in the list above. Let us first discuss means-end reasoning. According to our model of goal-processing, means-end reasoning is mainly performed after deliberation, so that it concerns chiefly chosen goals and future-directed intentions. This expresses the strong tendency of agents to reason about the proper means to achieve their intentions, whereas they do not necessarily feel compelled to analyze the means needed to bring about a mere desire, i.e. an active goal. As for the screen of admissibility effect, this is connected with the inertia characteristic of intentions, and these two issues are better discussed together.

Intentional inertia captures the key insight that, in resource-bounded agents as we are, deliberation has to be *settled* at some point, so that the agent can devote all energies to find a way of achieving what he had decided of pursuing (Bratman et al., 1988; Pollack, 1991; Cawsey et al., 1993; Bell, 1995; Wooldridge, Parsons, 1999; Schut et al., 2004; van der Hoek et al., this issue). If deliberation were an endless process, intentional action could never occur. This is the basic rationale for agents to have a peculiar resistance to drop or change their current intentions – and this

is also the ultimate reason why intentions act as filters for the acquisition of further incompatible intentions, since being faced with incompatible options would force the agent to resort again to deliberation, shelving for the time being the pursuit of a previously intended aim. For the same token, in our model an intention-to do (or better, the Int-Act of a future-directed intention) is put in agenda just to execute the corresponding action at the appropriate time, not to be reconsidered; it has been already decided, and resources have already been spent and invested for it. So it is stable also because there are ‘sunk costs’ (Arkes, Blumer, 1985) and already committed resources.

Of course, it is not rational to consider deliberation settled forever, and in fact intentional inertia is not an infinite quantity or a universal constraint. After all, other opportunities or dangers may become known to the agent at a later time, and this would call for in-depth reconsideration of a previously established course of action. Consequently, the inertia of intentions is something which is supposed to be valid *all other things being equal* – in particular, in the absence of new information such as to overrule past decisions, therefore forcing upon us rational reconsideration of our choices.

Given this understanding of intentional inertia, it is easy to see how this feature is embedded in our model of goal-processing. Take a set G^0 of active goals at time t_0 , and imagine that, after passing all the belief filters in the model, this results in a consistent set I^0 of future-directed intentions at time t_0 . If we now assume that nothing relevant changes, at time t_1 the only new active goals presented to the agent attention will be those (if any) which were generated as means to achieve one of the chosen goals in I^0 : as such, these new active goals are not in contrast with the agent’s intentions and do not require any revision of the agent’s past choices. The only cases in which further deliberation (between conflicting ends, not between alternative means) may be needed is when new active goals are generated, e.g. by contextual activation, or when one of the beliefs supporting a prior intention is invalidated. But these represent cases in which all other things are *not* equal, so that intention revision may be in fact the rational option for the agent to take.

One may argue that this form of intentional inertia is too weak, since it expresses only the tendency of an agent to focus on his intentions as long as no other option is made manifest. In contrast, we may be interested to model *stronger forms of inertia*, in which even new alternative options have to overcome an intrinsic primacy which is given to those goals that the agent had already chosen. This clearly amounts to discuss *what kind of commitment* an agent should have towards intentions. Our claim in this respect is twofold. On the one hand, we maintain that Bratman’s analysis of intentional inertia requires only a weak form of commitment, because the formation of an intention is supposed to settle deliberation concerning only *that* option with reference to a *pre-existing set of alternatives*, and not deliberation concerning other options and/or

with reference to a modified set of alternatives. On the other hand, we deem certainly useful for a theory of intention to be able to express different forms of intentional commitment, and we consider our model of goal-processing to be adequate in this respect. First of all, the growing structure of supporting beliefs that a goal acquires through its processing can be seen as (part of) an analysis of the growing commitment that the agent is building towards that intention. Once a given supporting belief structure is in place, and especially after the deliberation stage, the agent is no longer free to abandon that chosen goal without reason – on the contrary, at least one of the beliefs supporting that intention must first be invalidated. In addition, the constructive theory of intentions outlined here could in principle accommodate several extensions aimed at modelling different forms of commitment (including over-commitment): e.g., a bonus in terms of value could be given to goals that have been already chosen, or they could be included as a further belief filter at the deliberation stage, i.e. generating their own specific incompatibility beliefs, or sunk costs could be introduced, keeping track of those resources that have been already spent for and committed to a certain future task. All these devices could be easily coupled with the model of goal-processing presented in 6, and they would effectively serve the same function that is delegated to the compatibility filter and the filter override mechanism in the BDI architecture proposed by Bratman, Israel, and Pollack (1988).

According to our analysis so far, all crucial features of Bratman's theory of intention are accounted for by a proper goal-belief model: intentions, once conceived as double-faced attitudes (see 8) compounded of chosen goals with their characteristic structure of supporting beliefs, are shown to be conduct-controlling attitudes, endowed with a characteristic form of inertia, and projecting specific constraints on the reasoning processes of the agent. According to Bratman, these distinctive features of intentions should be closely linked to the paramount role they play in *planning* – so much so, that Bratman christened his own view as 'the planning theory of intention', to emphasize that intentions and plans must be understood within the same conceptual framework. We do not object to this insistence on planning, but we see no reason why reductionist analyses of intention should be unable to account for the role of intentions in planning – in fact, we repeatedly emphasized how planning is a necessary element in the transition from a future-directed intention on some end to a collection of present-directed intentions on its means. So our theory has no difficulty in showing how plans, understood as complex hierarchies of goals, can be «formed, retained, combined, constrained by other plans, filled in, modified, reconsidered, and so on» – exactly as auspicated by Bratman (1987: 8).

On the other hand, Bratman's view of the link between intentions and plans is arguably too strong, because it ignores the fact that plans are often *conceived prior to deliberation*, and therefore

not only as collections of intentions, but also as structures of goals, even before such goals are chosen as viable options to be pursued. Planning can be a thoroughly hypothetical activity, one that the agent performs in order to preliminarily check whether he disposes of the necessary means to achieve a specific desire (i.e. during the evaluation phase), or to the purpose of assessing more exactly the costs of pursuing a given course of action (i.e. at the deliberation stage). As Cohen and Levesque observe, «although there surely *is* a strong relationship between plans and intentions, agents may form plans that they never “adopt”, and thus the notion of a plan lack the characteristic commitment to action inherent in our commonsense understanding of intention» (1990: 215). So, while it is true that bringing about a future-directed intention necessarily requires some degree of planning, it is not true that planning is necessarily about intentions, contra Bratman. Agents can and do consider plans also directly at the level of goals that are not yet chosen; moreover, under the appropriate circumstances, they are rationally justified in doing so.

On a slightly different note, we have already emphasized in section 8 that there is a *kernel of instrumentality* necessarily embedded in the very definition of what an intention is: the Int-Act to do something is generated, maintained, and justified precisely as a means to achieve some Int-End, and it is the combination of these two chosen goals, through the link of a belief on instrumentality, that properly constitutes an intention. This implies that there is a planning dimension intrinsic to the nature of intentions, that is prior to and independent from any further planning that the agent may have to perform to realize his intended aim. This constitutive connection between intending and planning is different from the one stressed by Bratman: here the point is not that intentions are necessarily the ‘building blocks’ of plans, nor that «plans are intentions writ large» (1987: 8) – on the contrary, it is rather the case that *intentions are plans writ small*.

In conclusion, our main point is that intentions, although playing a crucial role in practical reasoning, can still be precisely *defined in terms of beliefs and goals* (a term that we prefer to the somehow confusing notion of desires; see 2). Therefore, we do not need to invoke intentions as a primitive notion, contra Bratman and contra all the BDI formalisms that were based on Bratman’s work. This conceptual analysis turns out to be crucial for formal models as well, precisely because here we are discussing *what should be the proper primitives to model intentional action*: which of them are truly necessary, and what is their expressive power. In fact, there are several reasons to uphold a constructive theory of intentions. With reference to formal models, the relevant question is: Whether and why should it be *better to have two primitives for mental states*, i.e. goals and beliefs, instead of three, as it is customary in BDI?

Aside from obvious advantages in terms of simplification of the formalism, and the possibility of including intentions as derivative notions (as it is the case also in the framework proposed by van

der Hoek et al. in this issue, where intentions are defined in terms of active plans and beliefs), we would like to propose the following reasons to prefer the goal-belief model of intention over Bratman's planning theory:

1. *Greater expressive power.*
2. Integrated analysis of both *similarities* and *differences* between pro-attitudes.
3. *Genetic* model of intentions.

Concerning expressive power, this is enhanced by a constructive theory of intentions in a variety of ways. First, the model of belief-based goal-processing presented in 6 immediately defines also an intermediate notion between desires (active goals) and intentions (double-layered structures of chosen goals), i.e. *pursuable goals*, or *candidate intentions*. At the same time, the model also accommodates a clear-cut distinction between future-directed and present-directed intentions, as discussed in 8. Finally, and most noticeably, this analysis specifies exactly *how a goal can become an intention* (in BDI terminology, what is that turns a desire into an intention), hence contributing to bridge the explanatory gap between volition and action. If we take intentions as primitive, as Bratman and the BDI theorists do, this crucial problem is by-passed rather than solved – and, in fact, to the best of our knowledge none of these authors concerned themselves with the issue of intention formation (on this point, see also Dignum and Conte, 1997).

Moreover, a crucial benefit of considering desires and intentions as goals at different stages of processing consists in being able to account not only for their obvious differences, but also for their (no less obvious) *similarities*. Indeed, it is quite surprising that such similarities are largely overlooked in the BDI framework, where nobody seems to heed the fact that desires and intentions are akin to each other in a way that beliefs are not. To start with, both desires and intentions are about how we would like the world to be, whereas beliefs concern reality as it is or will (probably) be. As a consequence, beliefs can be true or false, but this distinction does not apply to desires and intentions. Conversely, desires and intentions express a motivational urge that is absent in beliefs. Intentions, like desires, can be satisfied or frustrated, and they can be assessed in terms of success and failure, while this does not happen for beliefs. Desires and intentions are connected with specific feelings and emotions (e.g. hope, worry, regret) in a way in which beliefs are not. Moreover, both desires and intentions have the power to influence behaviour teleologically (the former potentially, the latter actually): beliefs, as we discussed in 1, are maps by which we steer, but do not provide us with any guidance concerning the aim of our actions – where we should go on the map, so to speak. For the same token, both desires and intentions can trigger planning and means-end reasoning, whereas beliefs only serve as building blocks in these reasoning process, but cannot prompt them. Finally, not only intentions can be challenged by other, incompatible

intentions, but they can also be compared with contrasting desires, for instance when Joseph's intention of marrying Mary is undermined by his desire of remaining single. All these manifest similarities between desires and intentions are somehow taken for granted in BDI models, but they are never explained: the habit of treating both desire and intention as independent primitive notions has the negative side-effect of hiding their common origin. In contrast, the goal-belief theory of intention makes this origin manifest, and the common features of desires (active goals) and intentions (chosen goals) are immediately explained by their being two different specifications of the same primitive notion – namely, goals.

Finally, it is worth considering that a constructive theory of intention, in contrast with approaches that take this notion to be primitive, adds another layer of analysis: not only intentions, as composite concepts, must satisfy their functional role in practical reasoning, but they must also do so by virtue of their constitutive elements – that is, there must be coherence between the rules governing the atoms (goals and beliefs) and the functional behaviour of the molecule (intention), and that behaviour must be showed to result from the interaction of those atoms. This projects a *genetic constraint* over the notion of intention, in addition to the *functional constraint* emphasized by Bratman and others. Whereas BDI architectures are concerned only with the latter, formalisms based on the goal-belief theory of intention must satisfy both constraints. This can make things harder on the formal side, but the reward consists in achieving a much deeper understanding of intentional behaviour. Besides, our analysis in this section was meant to show that intentions as compounds of chosen goals and instrumentality beliefs, once defined along the lines suggest here, effectively satisfy the standard desiderata for intention, so that in the end there are fair chances of getting a satisfactory match between functional and genetic constraints.

10. CONCLUSIONS: TOWARDS A CONSTRUCTIVE THEORY OF INTENTIONS?

In our view, the main contributions provided by our analysis can be summarized as follows:

- A *taxonomy* and a *dynamic model of the belief structure that supports our goals*, from their inception as mere desires to their realization through intentional action. This model is richer than existing architectures of the interaction between beliefs, goals, and actions, as we argued in 3.
- Some working hypotheses on the consequences of this structural interaction between goals, beliefs, and actions, with special reference to the *dynamics* of these notions. In particular, we emphasized the principles of cognitive integration typical of intelligent behaviour (see 4), and the role of belief tests in goal processing (see 6) and intention formation (see 8).

- An original conception of *intentions as double-faced teleological entities* (see 8), which was argued to be interestingly related to other important contributions in this area, including Sellars' observations on the link between intention-that and intention-to, Bratman's functional analysis of future-directed intentions, and the desire-belief theory of intentions endorsed, in several variants, by Anscombe, Davidson, Goldman, and Audi.
- A detailed discussion of the standing of our model with respect to other theories of intentional action, in particular *Bratman's planning theory of intention* (see 9).

Alongside with its merits, the current formulation of our theory shows also some *limits*, mainly connected with its preliminary nature and pre-formal status. Among others, we are particularly aware of the following open problems, that we intend to tackle in future work:

- The crucial notion of *commitment* is still merely mentioned and roughly sketched in our theory of intention formation, rather than properly analyzed. Although we are very much aware that intentional commitment cannot be reduced to doxastic conditions, the mechanism of the intentional agenda outlined in 7 remains mainly metaphorical and not yet operational.
- In this work we avoided addressing the problem of the *nature and/or status that goal-supporting beliefs should have*, in order to say that a given belief test in goal processing is passed, either through full scrutiny or by default. Some notion of implicitness is obviously required to account for this systematic filtering process, but what kind of implicit beliefs may be needed is not yet clear: beliefs as dispositions, or as implicit consequences of representations already explicitly endorsed, or as tacit representations, i.e. not under current attentive monitoring. This point will require additional clarification in future work, and a close comparison with the literature on other belief-like mental attitudes, such as assumptions, acceptances, presumptions, etc.
- The whole model of belief-based goal processing, although quite well specified in its details, is *not yet formalized or implemented*. Moreover, at present it remains unclear what could be the best way to formalize and implement it.
- *Connections with empirical data*, aside from anecdotic evidence and common-sense examples, are still *too scarce*. In future work, we intend to compare some of our ideas with empirical studies on intentional behaviour conducted in experimental psychology, developmental studies, and cognitive neurosciences.

From a methodological point of view, our analysis here remains informal and pre-formal – that is, we did not employ any formalism in describing our model, but strived nevertheless to provide clear-cut predictions and general postulates, that we certainly hope will prove useful for building better

formalisms of intentional action. The target of our reflections here has been the conceptual core of the theory of action, rather than its formal apparatus. In this respect, as we tried to make clear in our critical discussion of Bratman's work, we feel that some conceptual refinement is needed. Without denying to Bratman's planning theory of intention any of its numerous merits, we tried to show why and where this theory appears inadequate as a general frame of reference for formal models. At the same time, we suggest an alternative approach to purposive behaviour, one that does not take intention as a primitive notion, but rather analyzes how a mere goal can become a proper intention, by virtue of an appropriate structure of supporting beliefs. This proposal has the crucial advantage of accounting for the striking similarity between *desires* and *intentions*. The transformation of disordered and possibly inconsistent fancies into well-formed and coherent intentions is a kind of magic that we witness every day. This indicates a deep kinship that needs explaining, and we tried to illuminate it by means of the more general notion of goals. Hopefully, in due time this first conceptual outline might blossom into a full-blown *constructive theory of intentions*, i.e. a precise model of the genesis, functions, and revision of intentions.

ACKNOWLEDGEMENTS

This research was supported by the EU Project *MindRACES* within the Cognitive Systems area (contract number: FP6-511931), and by the PRIN 2005 Project *Information Dynamics in Knowledge Society*, co-funded by the MIUR and the University of Siena. We are grateful to our colleagues at the ISTC-CNR in Rome for many valuable discussions on goal dynamics and the nature of intentions, and to three anonymous reviewers for their extensive and insightful comments.

NOTES

¹ There are motivational processes that are *not* properly intentional (in the sense of 'being intended', not just 'being about'): e.g., merely reactive or instinctive actions, as well as unconscious motives. All these processes are certainly motivated, and usually endogenous, but they are not intentional in any strict sense. Here we focus on intentional action only, so we will confine our discussion to the role of beliefs in a specific kind of goal processing: *intention formation*, i.e. the process by which a goal becomes an intention.

² In previous work (Castelfranchi, 1997; Paglieri, 2006; Paglieri, Castelfranchi, 2006) we already discussed the converse pattern of interaction between goals and beliefs: the processes by which goals, desires, and motivations bias and shape belief formation and change.

³ For a different formal analysis of the connection between planning, beliefs, and means-end reasoning, see section 3.3 of van der Hoek et al., this issue.

⁴ Quite obviously, also means-end beliefs can act as motivating beliefs, i.e. inducing activation of adequate sub-goals to fulfil some higher end. However, for the sake of clarity, in what follows we shall classify means-end beliefs as pertaining to a later stage of goal processing. But let it be understood that, whenever a means-end belief is called upon to conceive a proper plan to pursue a chosen goal, this results in the activation of a corresponding sub-goal.

⁵ Here we endorse a strict and objective sense of *costs*, rather than their subjective or psychological interpretation. Every course of action and every cognitive process requires resources, and the very fact that those resources are allocated for that task prevents them from being used by other processes. But these costs are antecedent and partially independent from the psychological 'costs' (or sacrifices or losses) that we perceive when we are considering other alternative goals and decide to renounce to pursuing some of them.

⁶ This is a crucial node in goal processing, because it is here (and not before) that the agent, through proper deliberation, has to make sure he will choose a *consistent* set of goals to pursue. Up to this stage, consistency is not an issue: active goals can clearly be mutually contradictory, and the same applies also to candidate goals, because at that stage the agent has not yet decided which of these goals has to be pursued, therefore it is not yet committed to any of them. This reflects the well-known principle that desires, in contrast with intentions, can be inconsistent. More importantly, our model suggest an operational understanding of how an inconsistent set of desires can be ‘filtered’ into a consistent set of intentions – and what act as filters are, in our view, supporting beliefs.

⁷ The distinction between precondition beliefs and means-end beliefs may appear not very clear-cut, because both types of beliefs often share part of their *content*. Precisely because something is a necessary precondition for the action that realizes my goal (e.g., knowing the recipe for preparing the liver Venetian style), then achieving that something is also a means for pursuing my goal (e.g., learning such a recipe) – whereas the converse is not true, since certain means (e.g. instrumental actions) do not count as proper preconditions. In any case, the distinction remains relevant, because precondition beliefs and means-end beliefs have related but different *functions* in goal-processing. Given a chosen goal, if the agent does not believe the preconditions for execution of the corresponding action to be missing, then the goal can become directly executive, with no need for planning on means-end beliefs. On the contrary, if some belief on incompetence or lack of conditions is present, then a search for proper means is initiated, possibly resulting in further planning.

⁸ Incompatibility beliefs, if absent, let goal processing progress without any further deliberation, and in this sense they constitute a negative filter; if they are present, however, they are essential to trigger a test on preference beliefs, to assess which one of the alternative options is to be given priority by the agent.

⁹ We postpone to future work detailed analysis of what kind of predictions our model entails concerning the *destiny of a goal*, whenever one of its supporting beliefs is invalidated. However, it is intuitively clear that invalidation of different types of supporting beliefs (i.e., failure of goal processing at different stages) will be crucial in determining whether the goal is completely dropped, reconsidered anew, or merely suspended (see also section 4 on this point). On this ground, it shall be possible to develop a more principled account of the connections between (some instances of) belief revision and intention reconsideration, possibly leading to further expansion of existing formalisms, e.g. the framework presented by van der Hoek, Jamroga and Wooldridge in this issue (see in particular theorem 8 in their analysis).

¹⁰ Let it be understood that this characterization of different types of goal is *not* meant as a formal analysis (in contrast with Cohen and Levesque’s model, 1990), but rather as a summary of informal guidelines that, hopefully, might prove useful in inspiring formal models of the role of beliefs in goal processing. The development and critical discussion of formalisms based on this taxonomy is left to future work.

¹¹ We are grateful to Luca Tummolini for pointing to our attention potential connections between the notion of intentions in agenda and Castañeda’s theory of action (1975), later on criticized by Bratman (1999: 225-249). We are aware of both similarities and differences between our position and Castañeda’s analysis, but we leave to future work in-depth discussion of this topic – including also the related work by Shoham on intentional commitment as an obligation to oneself (1993).

¹² On this point, see also Bratman (1987: 37-41), and the distinction between weak beliefs and strong beliefs in section 3.5 of van der Hoek et al., this issue.

¹³ Cohen and Levesque champion a similar intuition (1990: 220), although their subsequent analysis differs sharply from the one presented here. In depth discussion of similarities and differences between our approach and their theory of rational balance is left to future work.

REFERENCES

- Anscombe, G. E. M. (1957). *Intention*. Ithaca: Cornell University Press.
- Arkes, H., Blumer, C. (1985). “The psychology of sunk cost”. *Organizational Behavior and Human Decision Process* 35, pp. 124-140.
- Audi, R. (1973). “Intending”. *Journal of Philosophy* 70, pp. 387-403.
- Bell, J. (1995). “Changing attitudes”. In: M. J. Wooldridge, N. R. Jennings (eds.), *Intelligent agents: ECAI-94 workshop on agent theories, architectures, and languages*. Berlin: Springer-Verlag, pp. 40-55.
- Bratman, M. (1987). *Intention, plans, and practical reason*. Cambridge: Harvard University Press.
- Bratman, M. (1999). *Faces of intention*. Cambridge: Cambridge University Press.
- Bratman, M., Israel, D., Pollack, M. (1988). “Plans and resource-bounded practical reasoning”. *Computational Intelligence* 4, pp. 349-355.
- Castañeda, H.-N. (1975). *Thinking and doing*. Dordrecht: Reidel.
- 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 V. Cantoni, V. Di Gesù, A. Setti, D. Tegolo (eds.), *Human & Machine Perception: Information Fusion*. New York: Plenum Press, pp. 235-254.
- Castelfranchi, C. (1998a). "To believe and to feel: The case of needs". In: L. Cañamero (ed.), *Proceedings of the AAAI Fall Symposium "Emotional and Intelligent: The Tangled Knot of Cognition"*. New York: AAAI Press, pp. 55-60.
- Castelfranchi, C. (1998b). "Modelling social action for AI agents". *Artificial Intelligence* 103, pp. 157-182.
- Cawsey, A., Galliers, J., Logan, B., Reece, S., Sparck Jones, K. (1993). "Revising beliefs and intentions: A unified framework for agent interaction". In A. Sloman, D. Hogg, G. Humphreys, A. Ramsay (eds.), *Proceedings of AISB '93*. Amsterdam: IOS Press, pp. 130-139.
- Cohen, P. R., Levesque, H. J. (1990). "Intention is choice with commitment". *Artificial Intelligence* 42, pp. 213-261.
- Davidson, D. (1980). *Essays on actions and events*. New York: Oxford University Press.
- Davis, W. (1984). "A causal theory of intending". *American Philosophical Quarterly* 21, pp. 43-54.
- Dignum, F., Conte, R. (1997). "Intentional agents and goal formation". In M. P. Singh, A. Rao, M. Wooldridge (eds.), *Proceedings of ATAL97*. Berlin: Springer, pp. 231-244.
- Georgeff, M., Lansky, A. L. (1987). "Reactive reasoning and planning". In: *Proceedings of the 6th National Conference on Artificial Intelligence (AAAI-87)*. Seattle: pp. 677-682.
- Goldman, A. (1970). *A theory of human action*. Englewood Cliffs: Prentice-Hall.
- Konolige, K., Pollack, M. (1993). "A representationalist theory of intention". In *Proceedings of IJCAI93*. Chambery: ACM Press, pp.390-395.
- Miceli, M., Castelfranchi, C. (2002). "The mind and the future: The (negative) power of expectations". *Theory & Psychology* 12, pp. 335-366.
- Miller, G., Galanter, E., Pribram, K. H. (1960). *Plans and the structure of behaviour*. New York: Holt, Rinehart & Winston.
- Paglieri, F. (2004). "Data-oriented belief revision: Towards a unified theory of epistemic processing". In E. Onaindia, S. Staab (eds.), *STAIRS 2004: Proceedings of the 2nd Starting AI Researchers' Symposium*. Amsterdam: IOS Press, pp. 179-190.
- Paglieri, F. (2006). *Belief dynamics: From formal models to cognitive architectures, and back again*. PhD dissertation, University of Siena.
- Paglieri, F., Castelfranchi, C. (2006). "The Toulmin Test: Framing argumentation within belief revision theories". In D. Hitchcock, B. Verheij (eds.), *Arguing on the Toulmin model*. Berlin: Springer, pp. 327-343.
- Poggi, I. (2005). "The goals of persuasion". *Pragmatics and Cognition* 13, pp. 297-336.
- Pollack, M. (1991). "Overloading intentions for efficient practical reasoning". *Noûs* 25, pp. 513-536.
- Ramsey, F. (1931). "Truth and probability". Reprinted in H. E. Kyburg, H. E. Smokler (eds.) (1964), *Studies in subjective probability*. New York: Wiley, pp. 61-92.
- Rao, A. S., Georgeff, M. (1991). "Modeling rational agents within a BDI-architecture". In J. Allen, R. Fikes, E. Sandewall (eds.), *Principles of knowledge representation and reasoning: Proceedings of the second international conference (KR91)*. San Mateo, CA: Morgan Kaufmann, pp. 463-484.
- Rosenblueth, A., Wiener, N., Bigelow, J. (1968). "Behaviour, purpose, and teleology". In: W. Buckley (ed.), *Modern systems research for the behavioural scientist*. Chicago: Aldine, pp. 368-372.
- Schut, M., Wooldridge, M., Parsons, S. (2004). "The theory and practice of intention reconsideration". *Journal of Experimental and Theoretical Artificial Intelligence* 16, pp. 261-293.
- Sellars, W. (1967). "Form and content in ethical theory". The Lindsay Lecture for 1967: Department of Philosophy, University of Kansas (on-line version: <http://www.ditext.com/sellars/fcet.html> - last consulted on 14/09/2006).
- Shoham, Y. (1993). "Agent-oriented programming". *Artificial Intelligence* 60, pp. 51-92.
- Singh, M. P., Asher, N. M. (1993). "A logic of intentions and beliefs". *Journal of Philosophical Logic* 22, pp. 513-544.
- Sycara, K., (1991). "Pursuing persuasive argumentation". In: *Symposium on Argumentation and Belief*. AAAI Spring Symposium Series, Stanford University.
- van der Hoek, W., Jamroga, W., Wooldridge, M. (2007). "Towards a theory of intention revision". *Shyntese*, this issue.
- van der Hoek, W., Wooldridge, M. J. (2003). "Towards a logic of rational agency". *Logic Journal of the IGPL* 11, pp. 133-157.
- Walton, D. N. (1998). *The new dialectic: Conversational contexts of argument*. Toronto: University of Toronto Press.
- Weiner, B. (1974). *Achievement motivation and attribution theory*. Morristown: General Learning Press.
- Wooldridge, M. J. (2000). *Reasoning about rational agents*. Cambridge: MIT Press.

Wooldridge, M., Parsons, S. (1998). "Intention reconsideration reconsidered". In J. P. Müller, A. S. Rao, M. S. Singh (eds.), *Intelligent agents V: Agent theories, architectures, and languages*. Berlin: Springer-Verlag, pp. 63-79.