Fabio Paglieri, Cristiano Castelfranchi

The Role of Beliefs in Goal Dynamics:
Prolegomena to a Constructive Theory of Intentions

Area 1: Agents and Mental Attitudes - ILIKS annual meeting - Trento, 30 November 2006

Overview of CM group work on mental attitudes

- Ongoing research topics in the Cognitive Modeling group at the ISTC-CNR (frequently in cooperation with the IRIT-CNRS):
  - Anticipatory mechanisms (Castelfranchi, Falcone, Pezzulo, Piunti, Tummolini)
  - Cognitive anatomy of emotions (Miceli, Castelfranchi)
  - Argumentation and belief change (Paglieri, Castelfranchi, Poggi)
  - Cognitive mediators of institutional dynamics (Tummolini, Castelfranchi, Conte)
  - Expectations, attempt, and surprise (Lorini, Miceli, Castelfranchi)
  - The evolution of cooperation (Tummolini, Paglieri, Conte)
  - Belief dynamics (Paglieri, Lorini, Pezzulo)
  - The nature and dynamics of trust and testimony (Falcone, Castelfranchi, Pezzulo, Paglieri)

Outline

- Some basics
  - Definition of goals
  - Principles of goal-belief coordination
- List of goal-supporting beliefs (tentative ontology)
- Dynamic model of belief-based goal processing
- Conceptual consequences for a theory of intentions
- Future work

Definition of goal

- 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*
- Defining function is to shape, to *direct* in a teleological sense the actual behaviour of the system: goal-oriented actions are actions *directed towards the realization of some specific state of the world*
- Cybernetic inspiration of this notion (Miller, Galanter, Pribram, Rosenblueth)

Principles of belief-goal coordination

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.*

BDI on belief-goal coordination

- *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 (Wooldridge, 2000: 1)*
- In BDI, this implies:
  - belief-based means-end coordination
  - belief-based commitment regulation (an agent cannot rationally intend *p* if it does believe *p* to be already the case, or if it believes *p* to be impossible)
- Part of our aim is to improve this oversimplified typology of goal-supporting beliefs

Goal-supporting beliefs: Tentative ontology

- *Motivating beliefs*: beliefs that activate certain goals
  - *Triggering beliefs*: beliefs that reactively activate goals on the basis of a pre-established association
  - *Conditional beliefs*: beliefs that activate a goal on the basis of the conditional nature of the goal itself
- *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
  - *Self-realization beliefs*
  - *Satisfaction beliefs*
  - *Impossibility beliefs*

Goal-supporting beliefs: Tentative ontology (ctd)

- *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
- *Incompatibility beliefs*: beliefs concerning different forms of incompatibility between different goals, that can force the agent to chose among them
  - *Terminal incompatibility*: goals cannot be both true in the same world (conflicting aims)
  - *Instrumental incompatibility*: goals cannot be achieved simultaneously (conflicting resources)
  - *Superfluity*: both goals are mere means to the same end (convergent means)

Goal-supporting beliefs: Tentative ontology (ctd)

- *Preference beliefs*: beliefs concerning what (incompatible) goals should be given precedence over others in the current context
  - *Value beliefs*, concerning the subjective value of a certain goal, given my current interests
  - *Urgency beliefs*, concerning when (if ever) a given goal will 'expire', i.e. it will be no more possible to achieve it
- *Precondition beliefs*: beliefs concerning the necessary preconditions for successfully pursuing a given goal by executing the appropriate action
  - *Incompetence beliefs*: beliefs of 'internal attribution', self-efficacy, and confidence
  - *Lack of conditions beliefs*: beliefs of 'external attribution', concerning external conditions, opportunities, and resources
- *Means-End beliefs*: beliefs on the instrumental relation between a given goal and an action or an event which is considered to serve to achieve the former

Belief-based goal processing

Goal Type	Process Stage	Supporting beliefs	Beliefs sub-classes	+/-
	ACTIVATION	Motivating beliefs	Triggering beliefs	+
			Conditional beliefs	+
Active Goals *(= desires)*				
	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*			

Inter-definability of different goal-types

- Active goal (desire): ***GOAL(p)***
- Pursuable goal: ***P-GOAL(p)*** =
  - *GOAL(p)*
  - **AND** *no assessment belief on p*
- Chosen goal (necessary for FDI): ***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*
- Executive goal (necessary for PDI)*:* ***E-GOAL(p)*** =
  - *C-GOAL(p)*
  - **AND** *no precondition belief on p*

Open issues with this model

- Is the order of different stages correct and/or rigid?
  - *Example: planning prior to deliberation*
- Are these processes sequential or parallel?
  - *HP: sequential reconstruction of the emergent dynamics of parallel processes*
- What kinds of beliefs do we have in mind?
  - *E.g. implicit vs. explicit beliefs*
- What about the role of non-doxastic factors in goal processing?
  - *E.g. goal activation through emotional arousal*
- Do we really need beliefs to account for certain stages?
  - *E.g. using preferences instead of value beliefs*

Contrast with Bratman on intentions

- According to Bratman, intentions are «distinctive states of mind, on a par with beliefs and desires» (1987: 20)
- In his analysis, as well as in BDI, *intentions are treated as a primitive notion*, in parallel with beliefs and desires
- Atomic view of intentions: intentions as mental atoms
- 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*
- Molecular view of intentions: intentions as mental molecules, formed by simpler atoms (i.e. goals and beliefs)
- *Methodological, non-eliminativist reduction*: (i) intentions do exist as specific and relevant mental states, that (ii) are formed by complex structures of simpler notions, i.e. goals and beliefs, so that (iii) their characteristic properties can be analyzed as an emergent effect
- Towards a constructive theory of intentions

The double-faced nature of intentions

- Before we claimed that an intention *requires* a goal at a specific stage of processing (i.e. a chosen goal)
- This is very different from saying that such a goal, in and by itself, *is* the corresponding intention
- Intentions are double-faced mental states: a chosen goal, i.e. a goal that we have elected among other to be pursued, immediately becomes a richer structure, 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)
- An intention is the *combination* of these two different teleological objects, and this double-faced structure is characteristic of intentions

The double-faced nature of intentions (ctd)

- Compare with Sellars on *intention-that* & *intention-to*:
  - Intentions are not limited to intentions *to do* (…). 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 "I shall do that which is necessary to make it the case that-p" (1967: 1-2).
- Every intention-that *entails*, *generates*, and remains *connected with* an intention-to as the vehicle for achieving the intended aim
- Intention-that and intention-to not as two different types of intentions, but as the *two necessary elements of any intention*

The double-faced nature of intentions (ctd)

- Whenever an agent has the *intention-that a certain result obtains* (Int-End), this necessarily requires a corresponding *intention-to do something* (possibly still unspecified) *in view of that end* (Int-Act)
- Both Int-Act and Int-End 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*

The double-faced nature of intentions (ctd)

- So, a future-directed intention on p requires:
  - a *chosen goal that p* (future-directed Int-End)
  - a *chosen goal of doing A* (future-directed Int-Act)
  - the *belief that doing A is a means to bring it about p*

To summarize:

- ***FDI(p)*** iff *C-GOAL(p) & C-GOAL(A) & Bel(A_means_for_p)*
- ***PDI(p)*** iff *C-GOAL(p) & E-GOAL(A) & Bel(A_means_for_p)*

Contrast with Bratman on plans & intentions

- Bratman's planning theory of intention, so called due to the strong link posited between intentions and plans
- According to Bratman, intentions are the building blocks of plans
- Conversely, «plans are intentions writ large» (1987: 8)

- Equally strong link between plans and intentions in our approach, but from a different perspective
- Although planning can also apply prior to deliberation (so that goals, rather than intentions, should be considered as the building blocks of plans), there is a kernel of instrumentality in the very definition of what an intention is
- So, we would rather say that intentions are plans writ small

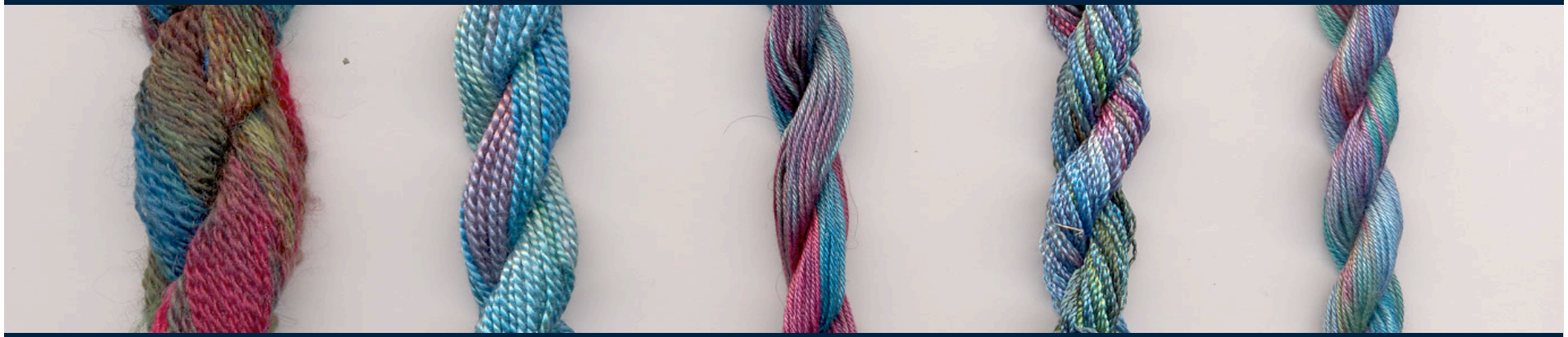
Constructivists do it better?

- A constructive theory accounts for Bratman's functional desiderata:
  - intentions are conduct-controlling attitudes, whereas desires (and goals) are merely potential influencers of action
  - intentions have a certain amount of inertia, so that agents show a characteristic resistance to drop or revise them
  - 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 *('screen of admissibility')*
- Additional advantages over Bratman's planning theory:
  - simplification of formalisms
  - possible to include intentions as derivative notions
  - greater expressive power
  - analysis of both similarities and differences between pro-attitudes
  - genetic model of intentions

Ongoing and future work

- Solving difficulties and ambiguities in our own model, as mentioned before
- Connection with the executive phase of intentional action
  - HP: intentions specify certain preconditions for the activation of an anticipatory classifier, which is the direct responsible for firing a given action
  - Here theory of intentions dovetails on analysis of simpler mechanisms for anticipatory conduct control
- The problem of commitment & intentional inertia
- Connection between belief change and intention revision (dependency corollary)
- Extend similar criticisms to Bratman's theory of action on the distinction belief vs. acceptance

Fabio Paglieri fabio.paglieri@istc.cnr.it
Cristiano Castelfranchi cristiano.castelfranchi@istc.cnr.it



Thanks for your kind attention.