Tirassa, M. (1997)
Mental states in communication.
Proceedings of the 2nd European Conference on Cognitive Science.
Manchester, UK, April 9-11, 1997.

Mental states in communication

Maurizio Tirassa

Università di Torino
Centro di Scienza Cognitiva
email tirassa@psych.unito.it

Abstract. This paper is concerned with the mental processes involved in intentional communication. I describe an agent's cognitive architecture as the set of cognitive dynamics (i.e., sequences of mental states with contents) she may entertain. I then describe intentional communication as one such specific dynamics, arguing against the prevailing view that communication consists in playing a role in a socially shared script. The cognitive capabilities needed for such dynamics are *mindreading* (i.e., the ability to reason upon another individual's mental states), and *communicative planning* (i.e., the ability to dynamically represent and act in a communicative situation).

INTRODUCTION

Agents are conscious organisms who live in dynamic situations, striving to make them more to their liking; a situation is an agent's subjective, open, changeable interpretation of the environment. Communication is a form of interaction between agents belonging to a socially sophisticated species; in particular, it is an overt attempt to reach a situation which be relatively satisfactory to all the participants.

The architecture of an agent's mind is the set of cognitive dynamics she[1] may entertain. I will describe cognitive dynamics as sequences of mental states, i.e., as abstract reformulations of folk-psychological descriptions like *Ann decided to go to the movies because she was bored to stay at home*. In this example, Ann's cognitive dynamics consists in a sequence of mental states (*decided*, *was bored*) with contents (*to go to the movies*, *to stay at home*). Mental states with contents, i.e., intentional states, are subjective, partial representations of the environment as it is, was, or might be.

An agent's cognitive dynamics must be tightly coupled to the partially unpredictable changes in her environment. I.e., they depend on the adaptive interaction between her cognitive system and the surrounding (mental, bodily, physical, and social) environment. The specific pattern of this dependence is rooted in turn in the evolutionary pressures which have shaped the agent's mind/brain and in her previous individual learning.

[1] Individual agents will be referred to with the feminine; in the case of communication, the masculine will be used for any agent involved but the first.

This paper is concerned with the specific cognitive dynamics underlying communication. Communication is a form of social activity overtly aimed at modifying a partner's cognitive dynamics. From a psychological viewpoint, it can be viewed as the interaction between the cognitive dynamics of the agents involved; or, in a one-sided perspective, between an agent's own cognitive dynamics and the dynamics she ascribes to her partners. A crucial role is thus played by *mindreading*, i.e., by each agent's capability of comprehending the other agent's cognitive dynamics.

I will match this view against the idea, prevailing in the relevant literature, that communication consists in taking up a role in a *joint plan*, i.e., a social script mutually known to the participants. I will argue that such idea is cognitively implausible, if only because it does not capture the flexibility typical of sophisticated sociality, instead allowing only for rigidly stereotyped patterns of interaction.

AUTONOMY AND SITUATIVITY IN COMMUNICATION

Although, after Grice (1989) and Searle (1969, 1979), communication is widely understood as involving some form of cooperation, little agreement can be found as to the nature of such cooperation. The more or less inordinate production of utterances is a form of cooperation in dialogue, which is satisfied when the agents succeed in making their communicative intentions mutually understood to the partners. Comprehension, however, is but the first step to affecting the partners' mental states in some specific way: if the agent's communicative intentions are indeed comprehended by the partners, but do not achieve the desired effects, they fail in spite of conversational coherence. In this respect, theories of communication belong to the study of social agency, rather than to linguistics or classical pragmatics.

Social interaction is often conceived of as the joint execution of a multiagent plan, mutually known to the agents involved (e.g., Airenti, Bara & Colombetti 1993; Grosz & Sidner 1990; Levesque, Cohen & Nunes 1990; Lochbaum 1993). The joint plan prescribes, to a variable degree of abstraction, a sequence of predefined actions to be performed by each participant; its execution is initiated by one agent and possibly carried on, upon recognition, by the others in their turn. Relevance is thus granted by each agent's adherence to the shared representation of the plan; describing a social interaction amounts to describing this plan and its conditions of applicability.

There are several problems with the joint-plan perspective:

- Like classical AI approaches to private action (which it straightforwardly applies to sociality), it cannot account for situativity. Plans, if conceived of as canned recipes to be traversed by an agent, can neither capture the uniqueness of each agent/situation coupling, nor adapt to its complex dynamics, nor account for the multiplicity and heterogeneity of an agent's goals (Agre & Chapman 1990). In the social case, this means that open systems (i.e., interactions among an unpredictable number of agents with heterogeneous, changeable goals: Gasser 1991) cannot be captured in a joint-plan framework. Turn-by-turn approaches to dialogue (e.g., Clark & Schaefer 1989; Sperber & Wilson 1986), on the other hand, may easily miss coherence, i.e., the (however relativized) persistence of an agent's intentions. Theories aiming at cognitive plausibility should capture an agent's long-term vision as well as her short-term flexibility.

- The indefinite number of canned recipes real agents would need gives rise to intolerable problems of learning, memory, and real-time reasoning. Further, the joint-plan perspective also requires of each participant to share exactly the same set of plans and of plan-recognition capabilities, in order to appropriately and timely recognize the particular plan the first agent is proposing to carry out.

- Built-in benevolence is another source of implausibility of the joint-plan framework. Behavioral cooperation is not an intrinsic feature of communication: only strictly task-oriented domains may be described as *usually* (but not *necessarily*) cooperative. Agents pursue their

own interests and participate in communication only as a means to carry them out; nor do they let them fall, once engaged in conversation, in favor of a superordinate aim to abstract cooperation. Cognitive theories of sociality must be neutral with regard to benevolence, antagonism, or indifference between agents (Castelfranchi 1990).

These difficulties may also be viewed as the symptoms of a more general problem: the map/territory fallacy, implicit in mainstream cognitive science, whereby behavioral regularities are unwarrantably explained away by postulating *ad hoc* rules encoded in the agents' cognitive machinery (Searle 1992). Cognition, on the contrary, is a biological phenomenon, and the mind/brain is not the implementation of a formal or computational system.

AN ARCHITECTURE FOR SOCIAL AGENCY

An agent is an intentional system who lives in a subjective, open, changeable interpretation of the environment (a *situation*), and strives to make it more to her liking. This definition is related, but not identical, to that of Pollock (1995); I have not the space to discuss the differences here.

Although an agent's knowledge can never be complete, she can improve it, within the limits of her architecture and of her current interests and resources, by zooming in (or out) on the environment, by reasoning from previous knowledge, and so on. Situations are thus changeable in accordance to the agent's knowledge and interests, as well as to perceived changes in the world. An agent's cognitive dynamics (i.e., the sequence of her mental states) must be coupled to the partially unpredictable changes in her situation; in the case of communication, situativity consists in a coupling between each agent's cognitive dynamics and the cognitive dynamics she ascribes to the partners.

At least six mental state types are needed in my account of communication. In the following, the notation S_a c reads *agent a entertains the mental state S with content c.* The content may also regard the very agent who entertains the mental state: thus, BEL_x happy(x) reads *agent x believes that agent x is happy.*

- *Beliefs* encode the agent's knowledge at a given time; their contents are situations and their characteristics. The notation BEL_x p expresses agent x's belief that situation p holds.

- *Mutual beliefs* are the social counterparts of beliefs. In Colombetti's (1993) definition, for agent x to have the mutual belief that p with agent(s) y is to have the belief that p *and* the belief that such belief is shared with agent(s) y: $MBEL_x$ (p, y) ≡ BEL_x p ∧ BEL_x $MBEL_y$ (p, x).

To act is to modify some characteristic of the environment, so to improve the current situation at a reasonable cost (in terms of difficulty, time, resources, etc.). To a deliberative (vs. purely reactive) agent, this means choosing within a (however limited) set of alternate futures. This is possible to an agent who is capable of representing some plausible, spontaneous or induced, evolutions of the current situation. I introduce three types of volitional states:

- *Desires* encode an agent's potential goals: $WANT_x$ α, where (α, β, …) are situations that the agent thinks she would prefer to the current one, independently of any further deliberation.

- *Future-directed intentions* (Bratman 1987) encode the agent's actual goals. Their contents are, among all the situations she desires, those she actually commits to: $FINT_x$ α', where (α', α'', …) are partial plans for situation α. A partial plan is a desired situation plus a course of possible activity, however ill-defined, which the agent thinks may achieve it.

- *Present-directed intentions* encode the agent's behavioral decisions. Their contents are basic actions: $PINT_X$ DO_X act1, where (act1, act2, …) are basic actions and DO_X act1 is agent x's execution of action act1. Basic actions are actions that the agent may immediately (i.e., with no further consideration) execute in the current situation.

An agent's volition is guided by her emotional/motivational system (Oatley 1992). Objects of thought, in other words, are never neutral; an agent assesses them as more or less satisfactory, and assigns to their components a causal role in her actual or potential overall satisfaction (Pollock 1995):

- *Likings* encode the agent's preferences between n-ples of situations, plans, or actions: $LIKE_X$ ($\alpha > \beta > \ldots$), $LIKE_X$ ($\alpha' > \alpha'' > \ldots$) and, respectively, $LIKE_X$ (act1 > act2 > …). Satisfaction is relative rather than absolute, as shown by the absence of absolute values in the notation.

AGENCY AND COMMUNICATION

Keeping in mind these mental state types, deliberation and private action may be described as follows. Given her current situation and motivation, the agent desires a certain set of situations, which she believes would be more satisfactory than the spontaneous evolution she forecasts of the current one. Because her resources are limited, and the world is partially unpredictable, she will not consider each and every possible/desirable situation: living in the real world neither consists in, nor allows, an exhaustive search in a closed space of predefined possibilities. Also, the alternate situations she forecasts or desires will be represented in poorer detail than the current one.

Since all of her desires can seldom come true, the agent will commit to a small subset of them, which she considers more satisfactory (or less costly) than the rest, and will make it the object of her future-directed intentions and further deliberation. E.g., in *Ann decided to go to the movies because she was bored to stay at home*, Ann's current situation may be described as *I'm getting bored*, her expectations about the spontaneous evolution of the situation as *I will get more bored*, and some desirable situations as *I would like to be in Morocco, I would like to see Bob*, and *I would like to go to the movies*; among these, she will have reasons to choose one (say, the latter) and commit to it. Because the world changes, and the agent has no complete knowledge of even its current state, plans can only be partial: they are better viewed as macro-actions to be further decomposed (*I'll go to the movies*), rather than recipes for action.

Plan execution starts with the generation of a present-directed intention to perform an appropriate basic action (*I take the newspaper to see what films are on, as part of my going to the movies*). Any action or event changes the world, opening some possibilities and closing others; it thus creates a new situation which can be more or less to the agent's expectations and likings. The agent then generates a new present-directed intention, consistent with her future-directed intentions and desires. These may have changed in the meantime, according to the contingencies and opportunities of the new situation, or as a consequence of a change in the agent's beliefs, motivations, or interests (e.g., Ann might find out, on reading the newspaper, that there is a beautiful concert on at the Music Hall).

Communication may be described in similar terms. An agent's actions in dialogue result from the interaction of her cognitive dynamics with those she ascribes to her partner(s). She will act, within the scope of her future-directed intentions, so to keep conversation up, until a situation is reached that she considers relatively satisfactory in terms of results, costs, possibilities left open, etc. The cognitive dynamics underlying communication consists in a turn-by-turn generation of suitable present-directed intentions to communicate, with feedback revision of the relevant desires and future-directed intentions. The main difference, with respect to private action, is that situation liking, and the alternate futures considered, now depend also on the cognitive dynamics the agent ascribes to her partner(s); specific mental states (such as mutual beliefs)

and knowledge are involved in this process, so that communication may be considered as a specific domain of cognition (Tirassa, in preparation).

Since the same description applies to each participant, communication may be viewed as the cooperative construction of a situation which be relatively satisfactory to all the agents involved. Each agent will keep communication up as long as she considers it useful and possible; i.e., as long as she believes that it may bring some worthy improvement in her situation, and that her partners are still willing to interact in their turn. Thus, the final situation will be satisfactory to all the participants. Satisfaction, to repeat, is relative rather than absolute: it depends on what the agent views as possible or worthy in the current situation. This explains, e.g., why soldiers seldom react to a sergeant's rebuke: silence, however unpleasant, brings less undesired consequences than rebellion.

Built-in cooperation is neither necessary nor desirable. Politeness, benevolence, and altruism may or may not be part of the agent's current motivations and interests: while social conventions prescribe that her participation in dialogue should not depend on her immediate urges alone, nothing impedes that she be rude (provided, of course, that she is ready to pay the relevant social cost).

AN EXAMPLE OF COMMUNICATION

This framework may be given a simple formal rewriting in terms of default dynamics of mental states, using the notation introduced in a previous section. I will here give a sketch of a proposal/reply exchange between agents x and y, from x's one-sided point of view. For reasons of brevity, I will describe a dual interaction, but the theory has no such restriction.

It is important to be clear about the role of the formalism. Consistently with my assumptions on the biological nature of cognition, I do not conceive of mental processes in terms of computations. Thus, the formal rewriting of a psychological theory is only meant to clarify and constrain the theory and possibly, when applied to closed situations, as might happen in a laboratory setting, to help explore specific predictions based on the theory (e.g. Bara, Tirassa & Zettin 1997). Formalism in psychology plays exactly the same role it plays in chemistry or in any other natural science. This is clearly a weaker position than is usual in cognitive science, but, in my view, one philosophically and psychologically more grounded.

Let us start with agent x's proposal:

[1]	$WANT_X\ \alpha$	; if agent x desires situation α
	$\wedge\ BEL_X\ ((\alpha', \alpha'', \ldots) \subset \alpha)$	; and believes that $(\alpha', \alpha'', \ldots)$ are suitable partial plans for α
	$\wedge\ LIKE_X\ (\alpha' > (\alpha'', \ldots))$	; and prefers α' to $(\alpha'', \ldots)$
	$\Rightarrow FINT_X\ \alpha'$	; then, by default, she future-intends α'

Default notation (Reiter 1980) is used because of the principled impossibility to list all the possible antecedents or consequents of a certain mental state; this may be viewed as a rewording of the frame problem.

E.g.: if Ann desires to spend the evening with Bob, and believes that two suitable partial plans to this aim are going to the restaurant with him and going to the movies with him, and she knows that there is a nice film on at the movie theater, then she may future-intend to propose that they go to the movies.

Then:

[2]	$FINT_X\ \alpha'$	; if x future-intends α'
	$\wedge\ BEL_X\ ((act1, act2, \ldots) \in \alpha')$	; and believes that (act1, act2, …) are basic actions for α'
	$\wedge\ LIKE_X\ (act1 > (act2, \ldots))$	; and prefers act1 to (act2, …)
	$\Rightarrow PINT_X\ DO_X\ act1$	; then, by default, she present-intends to do act1

A basic communicative action in a partial plan is a communicative action (i.e., a speech act) that the agent can execute with no further consideration, with the aim of making some fact mutually believed by her and y. E.g., Ann may decide between different speech acts to convey her proposal. From x's viewpoint, the result of executing act1 is $MBEL_x$ (DO_x act1, y); thus, Ann's proposal to Bob results in her mutual belief with him that she has executed it. From the given definition of mutual belief, this also implies that BEL_x $MBEL_y$ (DO_x act1, x).

If Ann believes that Bob has correctly reconstructed her cognitive dynamics, then:

[3] BEL_x $MBEL_y$ (DO_x act1, x) ; if x believes that y mutually believes with her
; that she has executed act1
$\Rightarrow \ldots \Rightarrow$ BEL_x $MBEL_y$ ($FINT_x$ α', x) ; then, by default, she believes that y mutually
; believes with her that she future-intends α'

i.e., she may believe that she and Bob now mutually believe *why* she has made her proposal.

Agent y has now to decide whether to comply with x's goal. At least, acceptance and rejection need be considered. From x's point of view, respectively:

[4a] $MBEL_x$ ($WANT_x$ α, y) ; if x mutually believes with y that she wants α
$\wedge$ $WANT_y$ α ; and that y himself wants α
$\wedge$ BEL_y ($\alpha' \subset \alpha$) ; and that y believes that α' is a partial plan for α
$\wedge$ $LIKE_y$ ($\alpha' > (\alpha'', \ldots)$) ; and that y likes α' more than (α'', …)
$\Rightarrow \ldots \Rightarrow$ $FINT_y$ α') ; then she may believe that, after further consideration,
; y future-intends α'

E.g., if Ann believes that Bob has understood that she wants to go to the movies with him, and that he finds this satisfactory, then she may believe that Bob too future-intends to go to the movies together.

As for rejection:

[4b] $MBEL_x$ ($WANT_x$ α, y) ; if x mutually believes with y that she wants α
$\wedge$ ($LIKE_y$ ($\alpha < (\beta, \ldots)$) ; but that y has different desires
$\vee$ BEL_y ($\alpha' \not\subset \alpha$) ; or y does not believe that α' is a partial plan for α
$\vee$ $LIKE_y$ ($\alpha' < (\alpha, \ldots)$)) ; or y prefers different partial plans for α
$\Rightarrow \ldots \Rightarrow$ $FINT_y$ $\sim\alpha'$) ; then she may believe that, after further
; consideration, y does *not* future-intend α'

e.g., if Ann believes that Bob has understood Ann's intention, but he prefers not to spend the evening with her, or he does not think that going to the movies is a suitable partial plan (say, because he believes that the movie theaters are closed that evening), or he prefers different ways to spend the evening together, then she may believe that, after further consideration, Bob does not future-intend to go to the movies with her.

In [3] and [4a,b], the expression "further consideration" (in the notation: $\Rightarrow \ldots \Rightarrow$) is used to short-circuit the modifications in y's desires, consequent to the modifications x has induced in his situation.

Steps [3] and [4a,b] are inferences that Ann can only draw after Bob's reply, i.e., they express her understanding of his reply. The cycle thus iterates. If Ann construes Bob's reply as a rejection, she will revise her future-directed intention or, if she considers this useless or unworthy, her desires. She might thus look for a different way to induce Bob to comply with her proposal, or for a different way to spend the evening with him; or she might give up her proposal and opt for the attempted achievement of a different situation (e.g., she might get rid of Bob and future-intend to go to the movies with someone else).

CONCLUSIONS

The framework for communication I have proposed builds upon a conception of cognitive architectures as possible sequences of mental states. From an abstract point of view, it has a single approach to private and social agency in terms of cognitive dynamics that are adaptively coupled to the environmental dynamics as known to the agent; situativity is thus an intrinsic feature of cognition, rather than a supplemental component to be added on top of an otherwise non-situated system. Situativity is made necessary by the rejection of the closed world assumption and by the impossibility for an agent to forecast and anticipate all the possible evolutions of the current situation.

A communicating agent's situation has to include the partner's (supposed) mental states, and her actions consist in speech acts. While it is impossible to share the details of each other's respective situation and cognitive dynamics, agents must share at least the basic elements of their respective architectures and be able to understand at least an outline of each other's cognitive dynamics. The potential complexity of an interaction is thus limited by the architectural differences between the participants, and by their capability to understand each other's cognitive dynamics, rather than by the number of social scripts they mutually know. Mindreading has thus a crucial importance in sophisticated sociality; indeed, it is a key architectural feature of a few species of highly social Primates (Byrne & Whiten 1988), whose disruption in autism (Baron-Cohen 1995) and schizophrenia (Frith 1992) hampers the development of any social interaction but the simplest.

Acknowledgments. This research was funded in part by the National Research Council of Italy (CNR), Coordinate Project on *Planning and plan recognition in communication*, 1995/1997.

REFERENCES

Agre, P.E., Chapman, D. (1990) What are plans for? *Robotics and Autonomous Systems* 6: 17-34.

Airenti, G., Bara, B.G., Colombetti, M. (1993) Conversation and behavior games in the pragmatics of dialogue. *Cognitive Science* 17: 197-256.

Bara, B.G., Tirassa, M., Zettin, M. (1997) Neuropragmatics: neuropsychological constraints on formal theories of dialogue. *Brain and Language*, in press.

Baron-Cohen, S. (1995) *Mindblindness. An essay on autism and Theory of Mind.* Cambridge, MA: MIT Press.

Bratman, M.E. (1987) *Intentions, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Byrne, R.W., Whiten, A., eds. (1988) *Machiavellian intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford: Clarendon Press.

Castelfranchi, C. (1991) No more cooperation, please! Controversial points about the social structure of verbal interaction. In: *AI and cognitive science perspectives on communication*, eds. A. Ortony, J. Slack & O. Stock. Heidelberg: Springer-Verlag.

Clark, H.H., Schaefer, E.F. (1989) Contributing to discourse. *Cognitive Science* 13: 259-294.

Colombetti, M. (1993) Formal semantics for mutual belief. *Artificial Intelligence* 62: 341-353.

Frith, C.D. (1992) *The cognitive neuropsychology of schizophrenia*. Hove, UK, and Hillsdale, NJ: Lawrence Erlbaum Associates.

Gasser, L. (1991) Social conceptions of knowledge and action: DAI foundations and open systems semantics. *Artificial Intelligence* 47: 107-138.

Grice, H.P. (1989) *Studies in the way of words*. Cambridge, MA, and London: Harvard University Press.

Grosz, B.J., Sidner, C.L. (1990) Plans for discourse. In: *Intentions in communication*, eds. P.R. Cohen, J. Morgan & M.E. Pollack. Cambridge, MA: MIT Press.

Levesque, H.J., Cohen, P.R., Nunes, J. (1991) On acting together. *Proc. 9th AAAI*, San Mateo, CA.

Lochbaum, K.E. (1993) *A collaborative planning approach to discourse understanding*. TR-20-93. Cambridge, MA: Harvard University.

Oatley, K. (1992) *Best laid schemes: the psychology of emotions*. Cambridge: Cambridge University Press.

Pollock, J.L. (1995) *Cognitive carpentry*. Cambridge, MA: MIT Press.

Reiter, R. (1980) A logic for default reasoning. *Artificial Intelligence* 13: 81-132.

Searle, J.R. (1969) *Speech acts: an essay in the philosophy of language*. London: Cambridge University Press.

Searle, J.R. (1979) *Expression and meaning.* Cambridge: Cambridge University Press.
Searle, J.R. (1992) *The rediscovery of the mind*. Cambridge, MA: MIT Press.
Sperber, D., Wilson, D. (1986) *Relevance. Communication and cognition*. Oxford: Blackwell.
Tirassa, M. (in preparation) Domain-specificity, mentalism, and cognitive pragmatics. PLEASE NOTE: THIS PAPER WAS LATER PUBLISHED AS: Tirassa, M. (1999) Communicative competence and the architecture of the mind/brain. *Brain and Language*, *68*, pp. 419-441.