



**HAL**  
open science

## Agency in schizophrenia from a control theory viewpoint

Joëlle Proust

► **To cite this version:**

Joëlle Proust. Agency in schizophrenia from a control theory viewpoint. N.Sebanz & W. Prinz. The disorders of volition, MIT Press, pp.87-118, 2006. <ijn\_00141931>

**HAL Id: ijn\_00141931**

**[https://hal.science/ijn\\_00141931v1](https://hal.science/ijn_00141931v1)**

Submitted on 16 Apr 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

## Agency in schizophrenia from a control theory viewpoint

**Joëlle Proust**

Institut Jean-Nicod, CNRS, Paris

1 b Avenue Lowendal

75007 Paris, France

tel : 33 1 53 59 32 87

fax : 33 1 53 59 32 90

email : [jproust@ehess.fr](mailto:jproust@ehess.fr)

number of words : 13198

number of figures : 0

total word equivalent

acknowledgment information

The research presented in this chapter has been supported by the European Science Foundation EUROCORES programme *The Origin of Man, Language and Languages*.

The author expresses her gratitude to Sliman Bensmaïa for his linguistic help.

### **Abstract**

Experience of agency in patients with schizophrenia involves an interesting dissociation; these patients demonstrate that one can have a thought or perform an action consciously (ie. have the sense of ownership for this action) without being conscious of thinking or acting as the motivated agent, author of that thought or of that action (without having the sense of agency associated with it). This chapter examines several interesting accounts of this dissociation, and aims at showing how they can be generalized to thought insertion phenomena. It is argued that control theory allows such a generalization; three different comparators need to be distinguished: the sense of subjectivity relies on a comparator in which motivation and emotion play a structuring role. The sense of agency emerges in a system that delivers a rough categorization of self-generated – versus other- generated –

actions and mental activities. A third system specializes in the social evaluation of the effects of an action, intention or other thought process, given certain goals in self or in others.

**5 keywords** to appear at the beginning of the chapter.

1. metacognition,
2. metarepresentation,
3. schizophrenia,
4. Sense of agency,
5. sense of ownership

**20 keywords** for the index.

1. comparator
2. contention scheduling system
3. control
4. delusion
5. efference copy
6. emotion
7. hallucination
8. hypnosis
9. intention
10. metacognition,
11. metarepresentation,
12. schizophrenia,
13. selective attention
14. Sense of agency,
15. Sense of ownership
16. simulation
17. social attribution
18. supervisory attentional system
19. thought insertion
20. working memory

## *Agency in schizophrenia from a control theory viewpoint*

Joëlle Proust

Institut Jean-Nicod

(CNRS, EHESS/ENS, Paris)

There does not seem to be a consensus as to the components of and processes underlying willful activity, or on the functional structures that are engaged in voluntary action. What makes the problem still more intractable is the difficulty, in the present state of the art, in offering an account of voluntary action that applies both to physical and to mental kinds of actions. In what sense imagining, remembering, or planning can be seen as voluntary mental actions rather than something that happens to the thinker? Is there a sense of agency that is common to physical actions, such as : opening a door, and mental actions such as : focussing on a problem?

The importance in studying impairments of willful activity lies in the fact that the scope of possible action-related states and feelings turns out to be wider than what our folk-psychological intuitions suggest. There is more to voluntary action than a simple “yes” or “no” answer to the question : « is this action *my* action ? » As many authors have observed, some subjects with schizophrenia, as well as brain-lesioned patients with alien hand syndroms, present a strange dissociation between the feeling that their own body is moving – an experience of *ownership* related to the fact that something happens to the self,- and the feeling that their body is moved by a foreign intention, rather than by the subject’s own will. For example, patients complain that their hands are moved by some irresistible external force. Although they do not acknowledge the action as theirs, they do identify it as their own. Thus *ownership* can be experienced while *agency* is not.

A fact that makes this dissociation all the more remarkable and relevant to the study of volitional states is that it extends to bodily as well as to mental actions. Patients who experience a lack of control on their bodily actions sometimes also have the feeling that their thoughts do not belong to them ; these are experienced as « inserted » in the patients’ heads, a sensation that is different from, although

related to, more classical forms of auditory-verbal hallucinations : inserted thoughts are perceived « internally », while auditory hallucinations are referred to an external speaker.

Some philosophers are insisting that there is nothing to be learned on normal function from psychopathology (Ricoeur, 1971). Some also maintain that patients with schizophrenia only display their irrationality in denying self-evident facts, of a kind that is taken to be « immune to error of misidentification » (Coliva, 2002). In other words, it seems impossible to believe that an action is performed by me, but not as a consequence of my intentions ; nor is it apparently possible to be mistaken as to who I am, because in both cases the thinker does not need to identify herself as an agent or as a self, but enjoys a form of direct, « non-observational » knowledge.

If it is recognized, however, that brain states and processes form a highly modular structure, evolved in steps, the view that the self offers a single unified model of the whole mind/brain's ongoing states tends to lose its appeal. A widely-held claim in cognitive science is that there are different levels of selfhood, from sensorimotor integration in an egocentric frame of reference, to more complex levels of self attribution in a social context (Rochat, 2003). The separation of these levels can benefit from the study of dissociations exhibited by lesioned or deluded patients. Specific functional disconnections are associated with phenomenological changes in patients' experiences of agency, which may help us expand our understanding of the dimensions of self-awareness. As many authors have observed, any functional change of this kind will retroact on other functions, bring new emotions into play, call for compensatory mechanisms, etc. But this may not prevent us from distinguishing crucial features reflected in patients' reports and cognitive performances

We will therefore focus on schizophrenic impairments of the sense of volition with a dual motivation : first, in order to better understand the dissociation between sense of ownership and sense of agency, a dissociation that defies folk-psychological intuitions; and second, to use this specific problem as an opportunity to scrutinize the notion that the mind is a set of nested control structures.

The structure of my argument will be as follows. I will first describe clinical facts related to the Ownership/Agency dissociation that need to be accounted for (section

1); I will then discuss the metarepresentational view of delusions of control as developed by Tim Shallice and Chris Frith in their classical studies (section 2). In Section 3, I will consider more recent views of the disorders of volition in schizophrenia. Section 4 will introduce proposals for relating the control & monitoring view of the mind/brain to metacognitive capacities engaged in volition. Finally, Section 5 will offer a revised picture of the impairments of volition in schizophrenia. According to the proposed hypothesis, three different control loops are engaged in the senses of ownership, of agency, and of explicit social attribution of actions.

#### 1- Four intriguing features of impaired will in patients with schizophrenia

##### 1.1. The ownership/agency asymmetry

We saw above that experience of agency in patients with schizophrenia involves a dissociation where none exists in a normal subject ; these patients demonstrate that one can have a thought or perform an action consciously – in the sense that they have the characteristic impression of having a thought or of executing an action - without being conscious of thinking or acting as the motivated agent, author of that thought or of that action. Phenomenology thus splits up into two different dimensions whose relationship is distinctively asymmetrical. *Whereas there is no case of an impression of agency without an impression of subjectivity, a sense of subjectivity can survive when the sense of agency is lost.* It is one of the aims of a proper theory of conscious experience to explain such an asymmetry.

##### 1.2. The parallel phenomena of thought insertion and delusion of control

Another challenge to a theory of volition (and of its disorders) has to do with its scope : is a single theory able to deal with willful thinking processes *and* willful bodily actions? It seems quite natural to require that a theory of volition provide a common theory of agency in both kinds of cases, as the phenomenologies in hallucinating patients with schizophrenia are very similar : an idea is entertained, an action is performed accompanied by a subjective impression, but both are sensed as having an externally generated, motivationally incongruent intentional content. This analogy has been spelled out either by taking thoughts to be actions of some sort, or by considering both

thought and action as involving a common metarepresentational format, which would be disrupted in schizophrenia. We will explore these avenues below, and provide a third explanation that builds on both ideas.

### 1.3. The external attribution puzzle

The puzzle can be summarized in this way: supposing that a patient with schizophrenia is impaired in monitoring her own intentions, actions, and thoughts, why does she not simply recognize that something is wrong with her ability to keep track of what she does and thinks? Why does she instead come up with odd judgments, such as that her neighbour, or some unknown person she met in the street, has taken control of her brain/body? What is the cognitive basis of «extraneity», one of the major symptoms in schizophrenia?

### 1.4. The occasionality problem.

This problem presents an additional difficulty to the preceding puzzle. Patients deny being the author of an action or of a thought only in certain cases; they seem to be able to have an irregular disposition to 'project' their mental contents onto others. The disposition is irregular in the sense that no general property of the projected content (such as : its emotional significance) seems to explain why the patient attributes it to another thinker or agent.

Our goal in this chapter will be to discuss accounts of schizophrenic cognitive impairments that lead to provide an integrated explanation of the first two features. Although the last two would well deserve an extensive discussion, it cannot be conducted within the confines of the present chapter.

## 2. Frith's metarepresentational view on self-monitoring

Chris Frith published in 1992 an influential book – *The Cognitive Neuropsychology of Schizophrenia* – in which the view that schizophrenia is essentially related to an impaired will was carefully presented and documented. Although Frith's theory was not meant to account for our first feature above (the asymmetry between sense of ownership and sense of agency), his 1992 theory contains the seeds of an explanation for it. There is an asymmetry between sense of ownership and sense of agency because first-order thoughts as well as routine

intentions and actions are preserved in patients; therefore, the phenomenology of perception and action is unchanged. Metarepresentations are impaired, however, which affects selectively the sense of agency as well as the explicit representations of the self. Let us first examine how volition is affected in schizophrenia according to this classical account.

## 2.1. Classes of volitional processes impaired in schizophrenia

According to Frith (1992), there are three major processes engaged in willful action that seem to be crucially involved in schizophrenic symptoms. A) The generation of intentions to act is massively impaired in patients who exhibit a «poverty of will»: patients with negative symptoms, in particular, may exhibit a reduced activity, a lack of persistence in their work, poor personal hygiene, and difficulties communicating with others. B) The monitoring of intentions is also often impaired: patients have difficulties selecting an appropriate action-schema; they also often have the feeling that the intentions driving their actions are not their own, and that their thoughts are inserted into their heads by other agents. The patients' impaired sense of agency seems to lead them to misattribute intentions to others: they may for example believe that other people are watching them (delusion of reference), plotting against them (delusion of persecution), or are having an emotional attachment to them (erotomania). In some cases, however, the patients attribute to themselves agency of others' actions. They feel responsible for other people's actions or even for large-scale world events, such as the war in Iraq. C) Finally, patients with schizophrenia monitor their actions in an abnormal way; they seem to be able to correct failed actions only if they have access to unambiguous visual feedback, in contrast to normal subjects, who seem to also rely on internal forms of monitoring (Frith & Done, 1989, Malenka et al., 1982).

In addition to these symptoms, which are directly involved in the sense of agency, two more symptoms have been mentioned by other authors as having an indirect relationship with action processes. One is the ability to *refer to self*, that seems disrupted in particular with respect to the use of personal pronouns such as «I» (A patient is reported to have told other patients in the ward «I am you, you and you» (pointing to three different individuals); the other is the related capacity to construct *an identical self* over time.

2.2. The metarepresentational theory of impaired intention, action and self monitoring.

Frith (1992) builds on Shallice (1988) to offer a simple explanation for the three main kinds of symptoms, which provides a parallel account for action and thought monitoring (our second feature above). Shallice's model for the control of action contrasts two functional levels; one is the «contention scheduling system» (CSS), which activates effectors on the basis of environmental affordances. It is taken to be a «low-level» system, that can perform routine or complex actions; it is regulated by mutual inhibition (winner takes all). However, according to this model, there is a higher-level form of control, called the «supervisory attentional system» (SAS). The latter is able to trigger non-routine actions, or actions that do not involve stimuli presently perceived. When SAS is active, it can invoke CSS-stored motor programs in an endogeneous way (action is no longer under the control of external stimuli). Various channels can be used to harness CSS programs to SAS, in particular natural language – which allows for the storage of plans of action and delayed commands in working memory - and episodic memory (in which a variety of situations are stored with their respective affordances).

Now, what is the functional difference between SAS and CSS ? Shallice hypothesizes that the Supervisory Attentional System has access to a representation of the environment and of the organism's intentions and cognitive capacities, whereas CSS only performs stimulus-driven, routine action programs. Thus the main feature of SAS that allows it to both provide an agent with a conscious access to her actions and to control routine actions is a metarepresentational capacity, that is, a capacity to represent oneself as having representations. An agent becomes able to act on her plans, instead of reacting to the environment, when she can form the conscious thought that she has such and such an intention.

Frith's 1992 theory works from Shallice's model to argue that an impaired metarepresentational capacity might account for distinctive features of patients' intentions and actions. «Specific features of schizophrenia, Frith writes, might arise from specific abnormalities in metarepresentation. This is the cognitive mechanism

that enables us to be aware of our goals, our intentions, and the intentions of other people.» If metarepresentation is disrupted, a patient will not only be unable to select actions endogeneously and to monitor them (for lack of a conscious representation of her own intentions), but also will be impaired in attributing an action or an intention to herself (or to others). Furthermore, impaired metarepresentation will disrupt conscious access to the contents of one's mental states. If metarepresentation is malfunctioning, an «imbalance» occurs between higher-level conscious processes and lower-level unconscious processes. As a result, patients become aware only of the contents of propositions, not of the metarepresentations in which they are embedded. Having had metarepresentations in the past, they are still able to attempt forming them. But they end up grasping only the embedded content: when trying to form the thought that someone thinks about P, they might only think «P». This same process would occur in inserted thought and in the sense of a loss of agency in action. Instead of considering some form of action, they will mistake the thought of a possible action for an order to act.

### 2.3. Discussion

This unifying theory, Frith (1992) admits, runs the risk of being «over-inclusive», in that it predicts that every form of metarepresentation should be the possible target of a symptom, whether in language, in social attribution, etc., which is not the case. On the contrary, Frith and his colleagues have observed that dissociations do occur in tasks supposed to tax metarepresentational capacity. For example, a patient can have trouble monitoring her intentions while being capable of inferring the intentions of others in indirect speech (Corcoran & Frith, 1995). Furthermore, patients with schizophrenia do not appear, as a rule, to be unable to report on their own mental states; rather, they are considered to be hyperreflexive (Sass & Parnas, 2001). An additional difficulty is that the model fails to account for the fact that a patient who has lost the sense of agency never admits that she does not know why she acts, but infers that *someone else* is acting through her own mind or body. Extranity remains a mysterious feature of schizophrenic experience.

3. Alternative accounts of the parallel between action control and thought insertion : the motor control view (Frith, Campbell, Jeannerod)

### 3.1 Frith's comparator model

In addition to the metarepresentational account summarized above, Chris Frith (1992) also sketches a motor-control explanation of delusions of agency, an explanation that turns out to be independent from the metarepresentational view and has since become the dominant view in the field, to the detriment of Frith's own former hypothesis. Patients' specific difficulty in monitoring their actions might be a consequence of a faulty or irregular efference mechanism of efference copy. In a normal subject, each time an action is launched, a copy of the intended movement is generated to compare it with the observed feedback. Such a comparator cuts down the amount of feedback required to check whether the action is successful, and makes control of action in normal subjects smooth and quick. In schizophrenia, the comparator might be faulty, thus depriving the agent both of the capacity to anticipate on the observed feedback and to consciously take responsibility for her actions.

How does this view deal with the parallel between action and thought ? Frith invokes Irwin Feinberg's idea (Feinberg, 1978) that thinking might also involve a «sense of effort and deliberate choice» : «If we found ourselves thinking without any awareness of the sense of effort that reflects central monitoring, we might well experience these thoughts as alien and, thus, as being inserted in our minds» (p. 81). It is not clear however in what a «sense of effort» might consist when no motor output is apparent. Furthermore, it is not clear whether the sense of effort presumably involved in thinking should be tagged as ownership (having a subjective feeling that one has a thought) rather than agency (the feeling that one is producing «deliberately» that thought»).

### 3.2. The Frith-Campbell view on agency in thought

In a series of papers (1998, 1999, 2000), John Campbell attempts to answer these two questions. According to him, the preserved sense of ownership in thought is dependent on what he calls «introspective knowledge», whereas the sense of agency in thought stems from a mechanism similar to efferent copy of action signals. Campbell hypothesizes that a *motor instruction* might normally mediate between background beliefs and desires, on the one hand, and the formation of a

thought, on the other: «the background beliefs and desires cause the motor instruction to be issued», which «causes the occurrent thought» (p. 617). This explains «how the ongoing stream of occurrent thoughts can be monitored and kept on track» (*ibid.*).

There are several problems raised by the Frith-Campbell's «control» model, in particular by the motor picture of thought formation. It has been observed that imagining performing action A, ( a particular kind of thought), activates motor-related brain structures that are normally active when A is performed. Some inserted thoughts have an imperative form, which might be interpreted as a failure to attribute intentions to self. But many cases of inserted thought do not include any reference to an action, and thus cannot be explained by misattributed agency. Why should motor activity be involved in thinking, for example, about the Pythagorean theorem ? It seems implausible, *prima facie*, to speculate that symbol activation and sentence generation «in the head» actually involve «manipulating» items, which would in turn require an efference copy mechanism.

A second objection is that many thoughts come to mind without a prior intention (or even without any "intention in action") that would make current ideation under immediate intentional control. Indeed if every thought presupposed a former intention, we would embark on an infinite regress. It does not seem that we normally intend to move from one thought to the next. The process of thinking does not seem to be constrained, in general, by former intentions. Furthermore, many of our ideas are not experienced as ours; for example, in a conversation, we process thoughts that are conveyed to us; we have no trouble both having the sense that we entertain a thought, understand it, process its consequences etc., and attributing its source to another thinker. The motor view, therefore, seems to deal rather poorly with the parallel between action and thought.

### 3. 3. Simulation and naked intentions

A different way of bringing action and thought closer is to consider action in its covert as well as overt aspects. This is the way in which Jeannerod & Pacherie (2004) approach the problem : they propose that the existence of overt behavior should not be a prerequisite for the sense of agency. They agree with Frith that the degree of mismatch between predicted and observed feedback modulates

activation in the right inferior parietal lobule and is responsible for external attributions of action (see also Jeannerod, this volume, & Farrer et al, 2003) ; the feeling of control is indeed inversely related to the activation level in this structure. However, they attempt to understand why covert actions, such as those performed in imagination, are also susceptible to be attributed to self or to other. Their solution consists in emphasizing two facts. First, simulatory mechanisms are elicited when observing as well as imagining and executing actions: intentions to act are thus represented impersonally, i.e. independently from the representation of an agent's having formed that intention and/or executed that action (such impersonal intentions are nicely labelled «naked intentions»). Second, patients with schizophrenia have a general difficulty in (covertly) simulating actions. Evidence from subjects with auditory hallucination suggests that they do not expect feedback from their own inner speech – another form of covert, simulatory activity. A defective simulation mechanism, rather than a defective action monitoring mechanism, could therefore be responsible for an impaired sense of agency in patients with schizophrenia. Unable to simulate the covert operations needed for attribution, these patients fail to identify some of their own actions and thus may misattribute them to others as well as misattribute others' actions to themselves.

Jeannerod & Pacherie's claim that there exists a neurally identified representational level that is common to overt and to covert behavior is clearly of major importance for understanding not only the nature of schizophrenic impairments, but also the sense of agency in normal subjects; their view implies that simulatory mechanisms have to be functional even before a comparator comes into play. Their account, however, raises several new and interesting questions. First and foremost, does impaired simulation *per se* underlie the patients' deficits ? How is it that a person with schizophrenia has no problem simulating *her own movements*, but has difficulty simulating goal-directed actions in the context of *attribution* ? Is self-identity affected by impaired self-attribution of actions, and how so ? How is the asymmetry between the sense of subjectivity and the sense of agency to be explained ? Finally, why is it that attribution of action and of thought are both impaired if the patients' primary disorder has to do with simulating bodily actions ?

In order to answer these questions, we need to be more explicit about the

architecture of representations of action, and to better understand their dynamic relations with self-other representations. The present proposal explores the possibility that there is a functional connection between thought and action, which explains why extraneity (delusions of control and of influence) applies to thoughts as well as to actions. This proposal is compatible with a view such as Jeannerod & Pacherie's, in which simulation is at the core of the generation and understanding of action.

#### 4 . An alternative proposal : clearing the theoretical background

We take the main contributions of Frith and Jeannerod's groups to be the following. Frith had two important but very different hypotheses. One was that very different symptoms in schizophrenia can be accounted for by an impaired metarepresentational capacity, whose links with executive competence have been emphasized by Shallice. The second was that patients with schizophrenia fail to identify and to self-attribute their intentions as a result of a failure in a comparator device. Jeannerod proposed that the main impairment is not related to the monitoring of action, but rather involves the simulatory mechanisms that make feedback prediction possible.

The present proposal aims to retain the advantages of a theory that accounts for a disturbance of thought as well as of action, as in Frith's metarepresentational view, while also trying to understand the specificity of the delusional experience in schizophrenia in control terms, in a way compatible with some aspects of Jeannerod and Pacherie's proposal. In short, we want to show that the parallel between thought insertion and impaired volition is grounded in the very control structure of mental activity, a structure that is common to thought and action. Our claim is built on three assumptions that need to be made explicit before we attempt to articulate our hypothesis on the phenomenology of volition in schizophrenia. First, simulation occurs in the brain as part of a control-monitoring sequence. Second, control levels are embedded, such that lower-level processing can be re-used (redescribed for other goals) at higher levels. Third, metacognition, rather than metarepresentation, is used in many contexts requiring insight about one's own competences and informational states. Recent results show that metacognition is

present in animals with no metarepresentational capacity.

#### 4.1. Ubiquity of simulation and mental architecture

Simulation occurs because the essential structure of the mind is a control structure. Simulating a sequence of action, or a sequence of external events, amounts to building a dynamic model of the internal or external environment. This dynamic model is constructed covertly (in an implicit, non conscious way) on the basis of prior learning. It helps produce internal feedback that provides detailed anticipations of reafferences in the absence of any actual engagement in the world. This simulatory process has been shown to be involved both in the production and observation of actions by others (Decety & Chaminade, 2003, Grezes et al., 2004).

Given that negative feedback control introduces intrinsic delays in the sensorimotor loop, a useful simulation must combine a feedforward dynamic model of an effector with an internal negative feedback loop to reduce such delays (Miall et al., 1993). Such a “Smith Predictor” is a two-loop structure: the inner loop provides a prediction of the outcome of each motor command sent to the effector, whereas the outer loop provides a prediction of the feedback synchronous with the actual feedback. Thus it is an internal feedback mechanism that operates in a feedforward mode. As has been shown by Miall et al (1993), this type of model might be multiply realized in the cerebellum and underly a number of predictions concerning the timing or the sensory reafference of a variety of action signals (Miall & Wolpert, 1996).

Now a consequence of the present proposal is that, as far as control is concerned, there is no real contrast between mental and bodily action. Indeed a familiar claim in metacognitive studies is that thought as well as bodily action can be controlled (Nelson & Narens, 1990). Metacognition refers to the kind of *knowledge* that a cognitive organism has of its own cognitive functioning, and to the various *processes* that are involved in controlling and monitoring its own informational states. From this perspective, there is no fundamental difference whether control applies to external or to internal actions. In both cases, the brain uses its own internal states and stored reafferences to simulate and regulate its own processes. In both cases, the brain must model dynamically one or several

sequences of potential activity, launch the execution of one of these sequences, - overtly or only covertly - and compare the observed to the predicted outcome. These forms of control indeed require the same kind of predicted and observed feedback; they all develop through the same kind of monitoring processes. If this view holds, it could help us understand impaired intentions, actions and occurrent thoughts in one fell swoop, in the spirit of Shallice (1988) and Frith (1992), although metarepresentation would no longer play a role in the new model; as will be shown below, the unifying feature of the various processes impaired in schizophrenia involves metacognition rather than metarepresentation.

It may be objected that bodily action brings with it rich reafferences about the world, (that may later be covertly simulated), while mental activity apparently does not. What are the reafferences, say, of searching one's memory, or appreciating whether a plan is feasible? Furthermore, overt action engages not only the brain, but the body, in particular the hands, limbs and posture. The feedback in bodily action thus seems to be inherently non-mental, but rather corporeal-environmental, in contrast to the kind of feedback one has in situations involving metamemory such as the feeling of knowing, or the « tip of the tongue » phenomenon, etc.. Is there a true distinction between these two types of feedback ?

Let us first observe that bodily experience plays a role in bodily action insofar as it represents an action in progress and related external events; *bodily sensations* help us recognize whether the action being performed is coherent with our intention; these sensations are useful only insofar as they are *reafferences*, i.e. perceptions that depend on former commands and are anticipated by internal feedback (conscious perceptual imagery and non-conscious dynamic models). The same structure operates in mental actions. For example, a subject who tries to remember a proper name expects that she will have a specific experience prior to any actual remembering; she will either have the feeling of knowing, or the feeling of not-knowing the name in question. In the first case, the distinctive feeling might be triggered by the propagation of activation in her memory network. In the second, the absence of this expected level of activation might trigger a distinct state. Insofar as it correlates with actual success, this feeling of knowing is obviously of crucial epistemic relevance: it allows the subject to be confident in her memory, and to launch a search only (or mostly) when potentially successful. But this epistemic

property may be based on a reafferent signal associated with the activation level of a specific structure.

Another sort of metacognitive feeling is the sense of something being “feasible by me”. When planning to execute a new task, e.g. lift a heavy object, one has to simulate how to grip it, etc. Predicted reafferences based on prior experience with similar tasks help the agent decide whether she is able to handle the task at hand.

To summarize: although the word « reafference » is usually applied to perception, it may have a broader applicability. The feeling of knowing, just as the feeling of being able, or other metacognitive feelings, can also be counted as a reafference insofar as i) it is collected on the basis of a prior command, ii) It is the dynamic properties of a specific brain/mental process, i.e. properties relevant for action, that are collected; iii) its function is to help the perceiving agent predict whether or not her action will be successful. If this is correct, there is no real contrast, from a control viewpoint, between « internal » and « external » reafferences. In both cases, the information collected represents an objective state of affairs : in the case of a bodily action, the expected/observed content of a perception e.g., « that some external goal state is (about to be) reached » ; in the case of a mental action, « that some internal goal state is (about to be) reached » , e.g., « that the retrieved name is correct». Obviously, modeling the external world and modeling internal processes have different functions. The former kind of information processing has to do with coping with a changing world, the second with the limitations of internal resources and with the feasibility of the operations required by various mental processes. But the forms of control might nevertheless be quite similar and involve in part the same anatomical structures. The idea that these two forms of control have the very same structure whether the reafferences are internal, proximal or distal can be illustrated with two recent results in neuroscience.

First, Miguel Nicolelis and his collaborators have shown that a monkey can learn to control a robot to which it is connected in a « closed-loop brain-machine-interface » (BMIc) - the robot is directly wired to neurons in the frontal and parietal areas (Nicolelis, 2000). Using visual feedback, monkeys are indeed able to reach and grasp objects using the robot’s arms, without moving their own limbs. As learning develops, a functional reorganization in cortical areas takes place : the

function of this reorganization is to incorporate the dynamic properties of the BMLc into sensory and motor cortical representations. The important point for our present discussion is that monkeys learn to reach and grasp virtual objects with a robot *in the absence of overt or covert arm movements*. They can also learn to move a robot arm separated from their own body. This « mechanical actuator » is acting out the subject's motor intentions in the absence of any proprioceptive feedback: vision provides all the necessary feedback for the cortical mapping of new commands. This fascinating experiment shows that there is more to external action than goal-directed behavior using one's body. It suggests a more general view of action in which an agent acts anytime she applies a command and monitoring sequence, in order to reach an outcome. Whether or not bodily movements are used in this sequence is a secondary matter.

A second example demonstrates human subjects' ability to use visual information to control brain activity. A study by Bock et al. (2003) has shown that human subjects are able to control regional brain activity using a form of feedback still further removed from ordinary perceptual reafferences, i.e. real-time fMRI "neuro-feedback". Subjects in this experiment were provided visual access (through a brain-computer interface, BCI) to the BOLD responses of two preselected brain areas (supplementary motor area, and parahippocampal place area), and were able within a few sessions to accomplish a given "mental" task, i.e. to reach a preselected (de)activation level in these areas. This study offers a very good example of how metacognitive activity is engaged in reaching preestablished goals – here, produce a given BOLD response – in a way that applies indifferently to various brain areas (whether purely informational, like memory, or executive, like motor areas). In this type of approach, the very distinction between bodily and mental action becomes murky.

These two types of studies suggest that, contrary to prima facie intuitions, the experimental setting of a BCI or BMLc is in some sense perfectly ecological: the brain indeed interfaces the present states of the world and the past interactions of the organism with it, in order to better cope with future situations. In such an interface, there is no « internal state » that is not meant to reflect external states ; and there is no perception of the environment that is not meant to potentially influence future actions and motivations ; command and monitoring are the two

dimensions of plasticity that the brain uses to adjust itself to the world dynamics. In all these cases, the controlling brain might be simulating its own processes in order to extract from the simulated dynamics the relevant potential reafferences and the final outcome of the projected action.

#### 4.2. Diversity of simulation levels: control hierarchy across dynamical units.

Let us then come back to our prior question: is simulation *per se* the site of the deficit in patients with schizophrenia, as is claimed by Jeannerod & Pacherie ? These authors defend their view by emphasizing that there is a specific level at which actions are represented (retrieved, recognized, identified, imagined, planned) that is independent of the process of social attribution (“who did it” ?). This representational level is engaged when covertly simulating actions, whether overt (in acting) or covert (in observing actions in others, or in planning and imagining, one’s own). What would be impaired in patients with schizophrenia is the very experience of agency, (rather than the level of conscious control of action). Patients would have difficulty attributing overt *or covert* actions to themselves or others due to an impaired ability to simulate actions in general. As Jeannerod further explains in the present volume, changes in patterns of cortical connectivity might disrupt either the networks mediating different representations, or the relative intensity of activation in the areas constituting these networks.

But simulating action by invoking various representations of action is not the only process involved in metacognitive evaluation and in attribution of action. There are various ways in which an action – or any dynamic event, internal or external - can be evaluated; accordingly there should be different functional control loops at play, interacting in a semi-hierarchical way. The diversity of control levels is a direct consequence of the fact that a dynamic, flexible control structure such as a human mind must monitor sequences developing on various time scales and dealing with various environmental properties. An event can be judged *for its immediate adequacy, for its instrumental adequacy, for its present and future social consequences, and finally for its resonance with long term values and life goals.* Corresponding to these various time scales and interests, we find a succession of embedded control-evaluative schemas: at the *motor level*, the move is judged as correct or incorrect at a postural-spatio-temporal level; at the *action output level*, the

successful completion of the intended action is evaluated; at the *agency level*, the action is attributed to its author, whether the self or another individual; at the *social level*, it is deemed whether the action, executed by self or by another, is compatible with a set of values that the agent is attempting to realize in her behavior, in line with her status, life projects, etc.

If there is such an embedding of control levels, as seems to be the case, then the notion of simulation becomes a generic term that might apply at each control level, but each time *in a different way*. Some forms of simulation might involve the effectors (muscles, joints, proximal limbs), some might involve perceptual memories of the external world, whether physical (object positions, events to be reached) or social (forming bonds, gaining influence, etc.). Finally emotional properties should play a crucial role in retrieving the scenarios involved in specific plans or social attributions.

Given the existence of such a control hierarchy; the question remains how the brain manages to regulate the informational flow across the various levels. For example, once an action is simulated, an additional task involves sending the relevant information to another structure in charge of executing the desired higher-level task: is this representation to be used as an internal command to act, as a piece of evidence to be factored into a prediction of what another person will do, as a cue to understand hidden goals and specific (true or false) beliefs, as a key move in a value-laden larger scheme ? These various ways of using a given motor representation may look similar. But evidence from evolutionary biology suggests that they are in fact different: each one functionally builds on the former; they are acquired successively in phylogeny and in ontogeny (at least in humans, the only beings who normally possess them all in adulthood). Thus a major task that the brain has to carry out is not only to simulate actions, but to dispatch the output of a simulation to the specific contextually adequate control center. This task typically requires metacognitive processing.

#### 4.3. From metarepresentation to metacognition

An important and distinctive feature of schizophrenic delusions, independently of their specific themes, is that patients fail to appreciate the extent to which their thoughts are deviant with respect to norms. Most deluded patients seem to lack

insight into how others will evaluate their utterances and react to them. This symptom has often been taken to involve an abnormal capacity to represent others' mental states. It might be tempting indeed to interpret this fact, as well as those discussed at the end of the preceding section, as supporting the metarepresentational view; the various capacities distinguished above may seem to differ from each other only in that some are invoked directly while others are imbedded in metarepresentations such as "He desires X to do Y", "she evaluates that X will help her get O", "she sees herself as being good at Ying" etc. In this section, we want to show why metacognitive processes can be understood in a way that does not necessarily involve such a metarepresentational format.

The functional difference between metacognition and metarepresentation is obviously of major relevance to our present discussion. Whereas metarepresentation is a theoretical (language-based) capacity for reporting (explicitly) on the contents of mental representations, metacognition is a practical (implicit) capacity for guiding mental activity. As we saw in section 4.1, metacognition includes all the processes through which primary mental functions are subjected to evaluation and control on the basis of their informational features. A key example of metacognition is metamemory, which determines what is accessible in memory, and triggers memory searches when appropriate (Koriat, 1993) ; another example is perceptual attention, which determines priorities and thresholds in information intake. The type of metacognitive competence involved in agency encompasses control, monitoring, and self-attribution of thought, of intention and of action (planning, for example, relies heavily on appreciating contextually one's own ability to perform various tasks).

In the Shallice-Frith model, summarized in 2.2, the metacognitive realization that an action is being performed or that an intention is being entertained by the self or another agent, is required to be in a specific metarepresentational format. It has recently been found, however, that this is not the case. Recent findings in animal cognition indeed suggest that animals without a theory of mind ( i.e. unable to metarepresent their own states in any explicit way), such as monkeys and dolphins, are able to evaluate their present ability to perform a particular task; they can judge which of two tasks (for example visual density discrimination tasks : Smith et al. 2003) they are more competent to perform. It is plausible that a more basic capacity

to mentally simulate a process (running a forward model to collect internal feedback) allows for an implicit form of (metacognitive) self-knowledge, which cannot be made explicit in non-verbal animals. Thus metarepresentational capacity might crucially depend on existing metacognitive skills for its normal functioning.

## 5. A control view of impaired volition in schizophrenia

It should be clear by now that the distinction between overt and covert simulation does not carry much functional weight. The core of the present proposal is that impaired volition should affect bodily as well as mental actions. As we saw in section 4.2., simulation is part of a control structure that also involves other processing steps. We now need to identify which processing steps are affected when a specific intention to act is “disowned” and thus misattributed. A way of addressing this question in control terms consists in trying to reconstruct the various control loops that are involved in the experience of agency and in the attribution of action.

### 5.1. Step 1: “Mineness” of phenomenal experience.

Step 1 involves a control loop that is responsible for an elementary feeling of “mineness” – ownership – experienced when a perception (or action) cycle is developing. This feeling corresponds to the sense of the current experience being one’s own. A perception cycle refers to the programming of a certain pattern of exploration of the world and the reafferences produced as a result, which in turn yield a new cycle. The sense of ownership associated with all forms of phenomenal consciousness is generally analysed as including two types of representations : i) that the perception (intention, memory) is about some event, and ii) that the perception (intention, memory) is mine. Should a given perceptual event be divided into what is seen, and the fact that it is seen by [me] ? What is often disregarded in such an analysis is the fact that the sense of ownership is generally *not* explicitly reflexive ; in Perry’s terms, (Perry, 2000) a conscious experience reflexively « concerns » a subject, but this relation of an experience – or of a thought –, to the thinker does not need to be explicitly represented as an “articulated” constituent.

The self does not need to be represented as such in this elementary form of experience – a form arguably present in ‘selfless’ animals. What then is the correct analysis of the sense of mineness associated with perception or thought in general ?

From a control theory viewpoint, what makes an experience mine is that it involves a set of (monitored) reafferences of a specific kind. A plausible hypothesis is that the feature of the reafferences that carries the implicit reflexive value of *mineness* is a specific *emotional* marker: only perception-cum-emotion can trigger the appropriate motivation to respond to the world in a self-relevant way without the need for a representation of self. In this light, the sense of subjectivity can be seen as a primitive metacognitive feeling, analogous to a feeling of knowing. This feeling applies to bodily states, as well as to thoughts and experiences. It allows the organism to distinguish, on the basis of the reafferences, what ‘concerns itself’, i.e. how the affordances present in the environment relate to its own fitness.

An alternative hypothesis is to take the sense of mineness as essentially associated with body-centered spatial content. This does not seem to be a solution, however, as it begs the question how bodily sensations feel like they are mine ; furthermore, the sense of mineness applies to thoughts as well as to bodily sensations ; it is not clear why the latter should be taken as more primitive or more substantial than the former.

To make the case that emotion is a crucial factor in mineness, let us articulate what conditions have to be present for an experience to be sensed as one’s own; they should include three types of clauses :

1. The experience has a certain phenomenological intensity and categorical content that covaries with a state of the world.
2. Its phenomenology refers to a specific source (usually a given modality) that carries metacognitive information about the experience.
3. The phenomenology includes a marker of emotional value to the organism itself. This emotional value is associated with specific motivations and dispositions to act.

These three properties of a phenomenological state help clarify the kind of reflexivity that the sense of mineness entails. Clause 1 relates to the representational aspect of a (monitored) reafference (experiences carry content, i.e. they have intentionality). Clause 2 articulates the metacognitive constituent implicit

in phenomenology (as Brentano observed, perceiving a red thing involves an awareness that this thing is seen rather than heard ; but this awareness is « implicit », so it can lead to further control operations without being explicit and metarepresentational). Clause 3 explains why a subject treats reflexivity as self-*concern*: emotional perspective stems directly from the innate capacity to evaluate perceptual input, and to exploit the affordances in goal-directed actions.

In bodily action, emotions are normally associated with perceptual and, in particular, to somatosensory and proprioceptive reafferences ; in thought, there might be at least two sources of emotional content. One is the content of thoughts, which is associated with prior emotional responses to external events ; the other is the metacognitive value of thoughts as reafferences : for example, subjects can be worried about (proud of, ashamed of, etc.) their cognitive performances.

It is important to observe that reafferences carry an emotional marker of mineness whether a willed action or a passive movement is performed. This observation leads us to make a distinction between the sense of agency and the sense of ownership, in contrast with Thomas Metzinger (2003 & this volume), who takes the former to be a subcategory of the latter. The fact that the two phenomenal dimensions of experience may coalesce should not lead one to ignore their functional difference (possibly associated with different phylogenetic origins). The very fact that the sense of ownership is unimpaired in patients with schizophrenia, whose sense of agency is impaired, might thus be accounted for by the fact that the sense of ownership belongs to a primary functional/phenomenal loop that feeds into higher-level loops, e.g. the agency loop. This question will be addressed below in section 5.2.4.

It is likely, however, that other pathologies might directly affect the sense of ownership, which should in turn disturb higher-level control loops; if the present hypothesis is on the right track, an impaired emotional system might fail to trigger a feeling of self-concern, which should in turn massively disturb the agency-control loop and the attribution-of-agency loop. This might be the case in patients with Cotard syndrome: Gerrans (1999) argues that the absence of affective processing in these patients might explain that perception and cognition have no emotionally significant bodily consequences, and thus are not accompanied by feelings of ownership. We might speculate that the first level in our control system – underlying

the sense of ownership –engages the primary somatosensory cortex, the cerebellum as well as the thalamus and the amygdala (the latter structures providing a sense of mineness to the associated sensory-motor feature detected by the former: Damasio, 1994, LeDoux, 1996, Leube et al., 2003, Ruby and Decety, 2003).

In brief: the sense of ownership is pre-reflexive. It does not take a metarepresentation and the resulting self-attribution for an experience to feel like mine. Indeed ownership is implicitly experienced in the subject's emotional reafferences when she has this perception, performs this action or has this thought.

## 5.2. Step two : Sense of agency

A sense of agency does not amount to sensing that one's body is moving or that thoughts are rolling in one's head : such a feeling belongs to the experience of ownership; the proper locus of the sense of agency is the feeling that the movement, or the thought process currently performed, are performed intentionally. The capacity to develop contrasting phenomenologies for passive and active movement is a major condition of survival. The metacognitive model sketched above provides an account of this capacity.

### 5.2.1 Challenging the classical "motor" view

According to the "motor theory" view presented in Section 3, the agent recognizes herself as acting when observed and predicted sensory reafferences are delivering congruent messages. This requires 1) that an efferent copy of a motor command has been used as input by a specific forward model to generate a state estimate; and 2), that the latter is used to compare in another forward output model the predicted with the observed sensory consequences of motor commands. In this section, what we will argue is that 1) such a comparison proceeds on the basis of specific instructions as to which sensory afferences to take into account. 2) These instructions already depend on a metacognitive understanding of the task being performed. The agent might thus metacognize her being active (in thought or in action) by relying on a task-specific signal selectively perturbed in patients with schizophrenia. This proposal is still speculative, but it seems compatible with

classical data, and indeed finds support in recent studies.

### 5.2.2. Deluding the motor system.

To introduce the proposal, I will discuss work by Sarah Blakemore and colleagues (Blakemore et al, 2003, Blakemore, 2003) in which an interesting parallel is developed between a hypnosis-induced and a schizophrenic form of delusion of passive movement. In order to understand what distinguishes self-generated from externally generated sensory events in normal subjects, Blakemore et al. used an experimental paradigm based on hypnotic suggestion. All subjects were hypnotized prior to test. In an Active Movement condition, subjects were instructed to raise their left arm. In the Passive Movement condition, subjects were told that their arm would be moved by a pulley (the pulley did not actually move). Highly hypnotizable subjects moved their arm as suggested, but reported no feeling of agency (they took their arm to be raised by the device). The cerebellum and the parietal operculum were found to be more active in the passive condition.

Interpreting these data in terms of the dual loop forward model discussed above (5.2.), Sarah Blakemore hypothesizes that the forward output model is specifically involved in the sense of agency, while the forward dynamic model regulating the movement itself is not (because the execution of the movement was unchanged, whether sensed as active or passive by the subjects). She thus explains agency in classical terms, through congruency with expected feedback and input cancellation. Lack of congruency (experienced in the Passive condition) would trigger cerebellar error-messages to the posterior parietal cortex on the right side, gradually promoting the impression that the action was under foreign control (Spence et al., 1997, Farrer et al, 2003).

Interestingly, Blakemore (2003) offers two other possible accounts of the sense of agency, in an attempt to explain why a lack of congruency should be felt by the hypnotized subjects: (the latter actually did send a motor command to the arm, which presumably triggered an efferent copy signal: so why is a non-congruency signal delivered ?). The first possibility is that there might be a frontal route to the sense of agency; referring to evidence that hypnotic suggestion results in an increase in rCBF in left frontal cortical areas, she speculates that hypnotic

suggestion might prevent motor intentions to reach the forward output model. The second explanation draws from work on attention. Attention to a particular stimulus is well known to enhance its sensory processing. Indeed it has been shown recently that preparing to attend to an anticipated stimulus can modulate activity in sensory brain areas *before* stimulus onset (Driver & Frith, 2000). The brain can covertly prepare a template of what it is supposed to detect by making appropriate “baseline shifts” in sensory activation. It is thus plausible that one of the main effects of hypnotic suggestion might consist in restricting the subject’s attention to part of the stimulus (its passive component). Given this top-down influence of attention, the comparator would ignore stimuli that have to do with willful activity. This shows that the comparator might not be the decisive structure for attributing the source of an action to self or other. The forward output model on which the comparator bases its predictions might already be preselected or filtered by attentional mechanisms.

Blakemore suggests that a similar mechanism might underlie delusions of control in schizophrenia. However, this mechanism needs to be spelled out in more detail. We will see that, upon closer scrutiny, the proposal appears to contradict the classical view on one main point.

If hypnotic suggestion prevents motor intentions from reaching the forward output model in normal subjects, does it help us understand why patients with schizophrenia are misattributing their actions to external agents ? Presumably, a feature common to Blakemore’s subjects and to patients with schizophrenia is that, in both cases, an instruction is sent to the cerebellum to ignore the fact that the movement is indeed voluntary; in other words, the cerebellum selects a normal feedforward dynamic model while failing to associate it with the appropriate feedforward output model, which triggers right parietal activation.

This account has to be refined, however, for in fact the induced delusions are different: patients with schizophrenia attribute their movements to an *external* agency, whereas hypnotized subjects are simply experiencing their movement as *passive*. Sarah Blakemore suggests that this difference might be explained by the contrast between being aware and being unaware of one’s intention: the patient with delusion of control would know her intention, then misattribute it to an external source for lack of congruence in the comparator; the hypnotized subject would not know her intention, and thus would not attribute it to an external agency, but rather

would feel the movement as passive. However ingenious and promising, this reasoning is incomplete: how does, in this account, a deluded patient come to be aware of her having an intention ( before she acts) ? This is a crucial question indeed, one that needs to be asked to fully understand the sense of agency. Only a theory that relies on the covert part of intentional action, such as Jeannerod and Pacherie's or the present proposal, can offer a non-circular answer to that question.

Let us now briefly explore the second route. Here, the idea is that both hypnotized subjects and patients with delusion of control fail to use their *attention* in a normal way. Focussing their attention on sensations associated with passive movement, they accordingly report having no sense of agency. But here again, we need to understand why patients *ignore* the willful component of the reafferences in the output model. It seems clear that the attentional account also jeopardizes the classical view, which is shown to be crucially incomplete: if the comparator is indeed under attentional influence, then it only plays a secondary role in the impairment of the sense of agency. What needs to be examined is no longer the comparator, but the attentional commands that are sent to it.

### 5.2.3. A metacognitive interpretation of the alternative accounts.

The two kinds of explanation sketched in Blakemore (2003) can easily be made to merge into one unifying metacognitive framework. Metacognition is the capacity to regulate brain activity on the basis of incoming internal and external information. Attention allocation and task selection are essential components of this regulatory system. The top-down effect of a frontal inhibitory command (preventing the relevant intention to influence the output model) can be redescribed in attentional terms. Under hypnotic suggestion, or in a schizophrenic delusion of control, attention can be restricted to the passive features of the executed movement. Whether the specific attentional baseline shift documented by Driver & Frith (2000) is under prefrontal control is not indicated. But other studies have emphasized that a system exerting a top-down influence on the selection for stimuli and responses involves the superior frontal cortex and its projections to the dorsal posterior parietal cortex. This system might integrate top-down and bottom up information, in order to form and update the "salience maps", i.e. determine which objects should be selected for recognition and action even before they are perceived. (Corbetta & Shulman,

2002). It is plausible to speculate that the attentional restriction in hypnotized subjects and in patients is triggered by biased or impaired frontal signals, respectively, which fail to induce correct predictions of sensory reafferences in the parietal area (in the output forward model).

This explanation is compatible with traditional views on impaired executive memory in schizophrenia: in patients with schizophrenia, it may be that insufficient gain is allotted to a course of mental action, namely an intention to act physically or to develop a train of thoughts. The inhibitory/gain component of selective attention, usually taken to belong to control processes (Umiltà & Stablum, 1998) and likely a prefrontal lobe function, has been shown to be impaired in patients with schizophrenia (Franzen & Ingvar, 1975, Goldman-Rakic, 1991). Patients have a general problem discriminating familiar from new information, and maintaining their goals in 'working memory', well-documented manifestations of prefrontal malfunction. In Goldman-Rakic's terms, the disorder comprises "a breakdown in the processes by which representational knowledge governs behavior".

More recent work lends additional support to the metacognitive component of impaired agency in schizophrenia. It has been shown that the rostral prefrontal cortex (Brodmann Area 10) might be involved in prospective memory when task coordination is required (Koechlin et al., 1999, Burgess et al. this volume). The evidence suggests that this capacity is impaired in patients with schizophrenia (Elvevag et al., 2003). More generally, the function of this area might be to produce, recall and evaluate internally generated information ; it would contribute to the establishment of a task set before actual task performance (Christoff et al., 2003). Liddle (this volume) also develops an approach compatible with our metacognitive hypothesis, observing an abnormal activation of a supramodal 'motivated attention system' in patients with schizophrenia.

These findings suggest the need to introduce a significant change in the comparator story. It is not so much that the deluded patient with schizophrenia has an impaired output forward model, as the classical view proposes. Rather, she is impaired in her ability to keep track of her intention to act over time, and to inform the comparator of that intention. Shifting the main causal role to the intention to act implicates the capacity of a patient to use implicit metacognitive goals to monitor

and further control her actions.

We are now in a position to account for the parallel between delusion of control in thought and in action. The problem common to both symptoms might not primarily consist in a poor simulating capacity (a cerebellar-parietal function), but rather involves initiating and maintaining the proper simulation over time (a prefrontal top-down influence on the cerebellar-parietal function). The sense of agency, i.e. the feeling associated with actively thinking and moving, might thus originate primarily in a prefrontal-parietal metacognitive refference rather than from a parietal-cerebellar one.

#### 5.2.4. The asymmetry between sense of ownership and sense of agency

In the revised picture, the asymmetry between sense of ownership and sense of agency is an automatic consequence of an idea introduced in § 4.2 of a *semi-hierarchy of controlled processes*. The basic idea is as follows: the control and monitoring system used to generate the sense of ownership is linked to the *hic et nunc* of perception and action in a given context; it feeds back into upper loops, such as the intentional agency or the social attribution loops, which have longer time spans, and whose content depends critically on the output of lower loops. Thus, although higher and lower loops can influence each other's activity (an example was offered above by baseline sensory shifts in attention), one can speculate that only lower loop output can be inherited by higher level representations. More concretely, if a perceptual or behavioral event is not treated as mine through relevant emotional marker(s), it should fail to trigger further processing by the brain. Indeed an organism deprived of the emotions associated with a "mine" feeling in a given context would not have the necessary motivation to carry out further processing.

#### 5.3. Step three - Explicit agency attribution : a social control level

Various authors, in particular philosophers, have based the feeling of agency in thought on the whole set of epistemic constraints that shape the attribution of agency. According to most, it involves accessing the *content* of the thoughts involved; some even require identifying the *kind of thought* being entertained

(believing, imagining, etc.). Much research devoted to the feeling of agency in action such as Daprati et al. (1997) and Jeannerod & Pacherie (2004), as well as Blakemore (2003) has tended to identify the feeling of agency in action as a full-blown attributional mechanism, through which the experience of action is explicitly referred to an author, whether the self or another person. One thus tends to blur the distinction between two types of attribution: The feeling of agency is the automatic, often implicit sense that an action was performed willfully; this feeling – or at least its functional equivalent – has to be present in one form or another in every behaving organism. The attribution of agency, on the other hand, is a process that requires the concept of an agent (as the source of an intentional action). The capacity to categorize goal directed-actions performed by others is a precursor of such an attribution. The animal has a primitive way of recognizing agency in others, which probably relies in part on the same neural structures involved in action. While this attributional capacity obviously could not evolve in beings with no sense of agency, there are organisms that have a feeling of agency but no attributional capacity, however primitive. An explicit attributional capacity might be necessary to identify stable agents in a social group, while a sense of agency only plays a role in an agent's own engagement in goal-directed behavior. How should we characterize, then, the functional difference between implicit (or action-level) and explicit (or social level) attribution of agency?

### 5.3. 1. Social-level attribution and self-identity

In Proust (2003), arguments were made in favor of the view that the kind of control and monitoring involved in representing oneself as a stable entity, responsible for her deeds, and permanently engaged in corrective metacognition, occurs at a level distinct both from control loops involved in ownership, and from mechanisms involved in the feeling of agency. A *third* level of control is needed to form dynamic models of one's own self as well as of others in a social group. In a nutshell, it is argued that such a representation emerges from a capacity to attribute and to keep track of one's long term goals and values, and to revise them when needed. The same capacity can also be used to simulate observed agents engaged in individual, competitive or cooperative tasks, with possibly conflicting intentions or selfish goals. Recent work in neuroimaging is indeed compatible with

this view (Decety & Chaminade, 2003, Grèzes et al., in press).

We saw above that, in schizophrenic delusions of control, the sense of agency is impaired, whereas the sense of ownership is spared. The clash between conflicting reafferences (the movement is mine but I did not execute it) is solved, at level 3, in the explicit attributional style that is characteristic of human cultures: the subject attributes her actions to another agent (or attributes other agent's actions to herself). She accordingly feels herself either diminished and enslaved to others' wills, or as amplified and extended to other agents. The emergent disorder of attribution is located at the social level; it results from a complex control structure (Adolphs, 2003), which obviously exerts a top-down influence on other loops by creating an expectation of extraneity in incoming reafferences. As was shown by Decety and Sommerville (2003), executive inhibition (i.e. possibly the right lateral prefrontal cortex) plays a major role in this control structure by suppressing the prepotent self-perspective in order to understand another person. This control structure for social attribution also involves external controllers: As emphasized by Frith (this volume), individual volition is also influenced at that level by exogeneous constraints imposed on her by social partners. The susceptibility of the social control structure to external influence might account for the involvement of highly diverse areas across tasks.

### 5.3.2 Simulating one's actions vs. simulating others'

The present account provides a tentative answer to a question that we raised in section 3.3: why does a person with schizophrenia have a specific difficulty simulating goal-directed actions in the context of *attribution*, rather than *execution*? In the present analysis, the two capacities are functionally distinct. Simulating one's actions, or simulating others', rely in part on the same structures. But using a simulation to explicitly attribute an observed action to an agent requires a higher-order control loop, which integrates self representations as well as motivational and emotional judgments. Given that this control loop requires the inhibition of self-simulatory processes and disengagement from routine evaluations (a "prefrontal" capacity), it is plausible that patients with schizophrenia fail to attribute actions in contexts which require adjustment and fine-tuning. This explains why patients are much better in executing than in attributing actions.

## **Conclusion**

The present proposal is an attempt to account for the fact that patients suffering from delusions of control exhibit an impaired sense of agency while their sense of subjectivity both in thought and in action remains intact. The theory draws from studies showing the wide range of control structures in the mind/brain, as well as the important metacognitive apparatus through which these structures operate. It was suggested that three different comparators need to be distinguished: the sense of subjectivity relies on a "local" comparator. Motivation and emotion play a structuring role in the "mineness" of the reafferences collected by this comparator. The sense of agency emerges in a different system: cerebellar, parietal and prefrontal structures deliver a rough categorization of self-generated – as opposed to other-generated – actions and mental activities, a competence closely related to source judgment, present in many non-human animals. A third system specializes in the social evaluation of the effects of an action, intention or other thought process, given certain goals in self or in others. It seems to be present only in humans (and maybe in other primates and dolphins). It involves many structures, in particular the limbic system and the orbitofrontal and medial frontal lobe. The full-blown « human » understanding of agency results from our capacity to plan our mental and physical engagements, a second order kind of control structure that reinterprets the output of the feeling of agency.

This type of control is necessarily exerted on local control structures responsible for the sense of subjectivity. The resulting experience normally fuses agency, ownership and attribution in one single stream. It is not necessary, however, for a system endowed with a sense of subjectivity, to maintain a sense of agency or an explicit social attribution capacity; when the higher-level control structure breaks down, the lower level can still persist although the phenomenology changes as it operates in an uncontrolled (or abnormally-controlled) mode.

There may be different grades of loss of higher-level control, from cases of patients with schizophrenia whose moderate executive difficulties translate into impressions of occasional xenopathy, to patients with severe forms of dementia, who have lost any capacity to act autonomously. From this perspective, action and thought are similarly organized as controlled processes, and there is no a priori reason to treat

them separately: the solution offered here works in the same way for inserted thoughts and xenopathic actions.

Clearly, this proposal opens up new questions, whose relevance goes beyond purely theoretical considerations: What are the relative contributions of innate and acquired factors in the attributional system ? In particular, what is the impact of exogeneous social demands on attribution of agency and controlled action ? How can the motivational top-down influence of the social level on the sense of agency be functionally understood? How does a theory of mind interact with this system? Does theory of mind form an essential part of the human attributional system, or does it build upon it ? And finally, how, more generally, does (implicit) human metacognition benefit from an (explicit) attribution of agency ? Answering these questions will not only shape social brain science, it will also deeply influence social development and education.

## References

- Adolphs, R. (2003). Cognitive neuroscience of human social behavior, *Nature Review Neuroscience*, 4, 165-178.
- Benton, *Frontal Lobe Function and Dysfunction*, Oxford, Oxford University Press, 125-138.
- Blakemore, S.-J. & Decety, J. (2001). From the perception of action to the understanding of intention, *Nature Review Neuroscience*, 2, 561-7.
- Blakemore, S. (2003). Deluding the motor system, *Consciousness and Cognition*, 12,4:647-655.
- Blakemore, S.-J., Rees, G. & Frith, C.D. (1998). How do we predict the consequences of our actions? A functional imaging study, *Neuropsychologia*.36 (6), 521-529.
- Blakemore, S.-J., Smith, J., Steel, R., Johnstone, E.C. & Frith, C.D., (2000). The Perception of self-produced sensory stimuli inpatients with auditory hallucinations and passivity experiences: evidence for a breakdown in self-monitoring. *Psychological Medicine*, 30, 1131-1139.
- Bock, S.W., Weiskopf, N., Scharnowski, F., Mathiak, K., Goebel, R. & Birbaumer N., *Differential neuro-Feedback using a Brain-Computer Interface (BCI) Based on Real-time fMRI*, Posted Communication, Meeting of the European Society for Cognitive

Science, Osnabruck, 2003.

Campbell, J. (1998). Le modèle de la schizophrénie de Christopher Frith, in H. Grivois & J. Proust (eds.), *Subjectivité et conscience d'agir, Approches cognitive et clinique de la psychose*, 99-113.

Campbell, J. (1999). Schizophrenia, the space of reasons, and thinking as a motor process, *The Monist*, vol. 82, 4, 609-625.

Campbell, J., (2002). The ownership of thoughts. *Philosophy Psychiatry & Psychology*. 9.1:35-39

Coliva, A. (2002). Thought insertion and immunity to error through misidentification. *Philosophy Psychiatry & Psychology* Volume 9, Number 1.

Conant, R.C. and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system, *International Journal of Systems Science*, 1, 2, 89-97.

Corbetta, M. & Shulman, G.L. (2002). Control of goal-directed and stimulus-driven attention in the brain, *Nature Reviews Neuroscience*, Vol.31, March 2002, 201-215.

Corcoran, R., Mercer, G., Frith, C.D. (1995). Schizophrenia, symptomatology and social inference : Investigating "theory of mind" in people with schizophrenia. *Schizophrenia Research*, 17, 5-13.

Currie, G.& Ravenscroft, I. (2002). *Recreative Minds*, Oxford, Oxford University Press.

Damasio, A. 1994. *Descartes'Error*, New York, Harper Collins.

Daprati, E., Franck, N., Georgieff, N., Proust, J., Pacherie, E., Dalery, J. & Jeannerod, M. (1997). Looking for the agent, an investigation into self-consciousness and consciousness of the action in patients with schizophrenia, *Cognition*. Vol. 65, pp. 71- 86.

Decety J. & Chaminade, T. (2003). Neural correlates of feeling sympathy. *Neuropsychologia*, 41, 127-138.

Decety J. & Sommerville, J.A. (2003). Shared representations between self and other: a social cognitive view, *Trends in Cognitive Science*, 7,12, 527-533.

Desmurget, M. & Grafton, S. (2000). Forward modeling allows feedback control for fast reaching movements. *Trends in Cognitive Science*, 4,11, 423-31.

Driver, J. & Frith, C.D. (2000). Shifting baselines in attention research. *Nature Reviews Neuroscience*. 1 (2), 147-8.

Ellevåg, B., Maylor, E.A. & Gilbert, A.L. (2003). Habitual prospective memory in schizophrenia, *BMC Psychiatry*, 3,1/9

- Farrer, C. & Frith, C.D.(2002). Experiencing oneself vs another person as being the cause of an action : the neural correlates of the experience of agency. *NeuroImage* 15 :596-603.
- Farrer, C., Franck, N., Georgieff, N., Frith C.D., Decety, J., & Jeannerod, M. (2003). Modulating the experience of agency : a positron emission tomography study. *NeuroImage*, 18: 324-33.
- Feinberg, I. (1978). Efference copy and corollary discharge : implications for thinking and its disorders, *Schizophrenia Bulletin*, 4, 636-640.
- Fourneret, P. & Jeannerod, M. (1998). Limited Conscious monitoring of motor performance in normal subjects, *Neuropsychologia*, 36, 11, 1133-1140.
- Franzen, G. & Ingvar, D.H. (1975). Absence of activation in frontal structures during psychological testing of chronic schizophrenics, *Journal of Neurological and Neurosurgical psychiatry*, 38: 1027-1032.
- Frith C.D. (1992). *The cognitive Neuropsychology of Schizophrenia*, Hillsdale, Lawrence Erlbaum Associates.
- Frith, C. (1994). Theory of Mind in Schizophrenia, in (A. David ed.), *The Neuropsychology of Schizophrenia*, Hillsdale, Lawrence Erlbaum, 147-161.
- Frith, C. D., & Done, D. J. (1989) Experiences of alien control in schizophrenia reflect a disorder of central monitoring of action. *Psychological Medicine*, 19, 353-363.
- Frith, C.D., Blakemore, S.-J., & Wolpert, D.M. (2000). Explaining the symptoms of schizophrenia : Abnormalities in the awareness of action, *Brain Research Reviews*, 31, 357-363.
- Gallagher, S. (2000). Self reference and schizophrenia, in D. Zahavi (ed.), *Exploring the self*, Amsterdam, John Benjamins, 203-239.
- Goldman-Rakic, P. (1991). Prefrontal Cortical dysfunction in Schizophrenia: The relevance of Working memory, in B.J. Carroll, & J.E. Barrett, (eds.), *Psychopathology and the Brain*, New York, Raven Press, 1-23.
- Grèzes, J. Frith, C.D. & Passingham, R.E. (in press). Inferring false beliefs from the actions of oneself and others: an fMRI study. *NeuroImage*.
- Hoffman, R. (1986). Verbal hallucinations and language production processes in schizophrenia, *Behavioral and Brain Sciences*, 9: 503-517.
- Jeannerod, M. & Pacherie, E. (2004). Agency, Simulation and Self-identification. *Mind & Language*, 19, 2: 113-146.

- Jeannerod, M. (1999). To act or not to act, perspectives on the representation of actions. *Quarterly Journal of Experimental Psychology*.52A :1-29.
- Koechlin, E., Basso, G., Pietrini, P., Panzer, S. & Grafman, J. (1999). The role of the anterior prefrontal cortex in human cognition. *Nature*, 13 May 1999, 399:148-151.
- Koechlin, E., Corrado, G., Pietrini, P. & Grafman, J.(2000). Dissociating the role of the medial and lateral anterior prefrontal cortex in human planning, *PNAS*, 97, 13:7651-56.
- Koriat, A. (1993). How do we know that we know ? The accessibility model of the feeling of knowing. *Psychological Review*, 100, 609-639.
- Kristoff, K., Geddes, L.P.T., Ream, J.M. & Gabrieli, J.D.E. (2003). Evaluating self-generated information: Anterior Prefrontal Contributions to Human Cognition. *Behavioral Neuroscience*, 117, 6, 1161-1168.
- LeDoux, J. (1996). *The Emotional Brain*, New York, Simon & Schuster.
- Leube, D.T., Knoblich, G., Erb, M., Grodd, W., Bartels, M. & Kircher, T.J., (2003). The neural correlated of perceiving one's own movements, *Neuroimage*, 20: 2084-90.
- Malenka, R.C., Angel, R.W., Hampton, B., Berger, P.A. (1982). Impaired central error-correcting behavior in schizophrenia, *Archives of General Psychiatry*, 39, 101-107.
- Maruff, P., Wilson, P. & Currie, J. (2003). Abnormalities of motor imagery associated with somatic passivity phenomena in schizophrenia, *Schizophrenia Research*, 60, 229-238.
- Metzinger, T. (2003). *Being No One*, Cambridge: MIT Press.
- Miall, R.C., Weir, D. J., Wolpert, D. M. & Stein J.F. (1993). Is the Cerebellum a Smith Predictor? *Journal of Motor Behavior*, 1993, Vol 25, No.3, 203-216
- Mlakar, J., Jensterle, J. & Frith, C.D. (1994). Central monitoring deficiency and schizophrenic symptoms, *Psychological Medicine*, 24, 557-564.
- Nelson, T.O. & Narens, L. (1992). Metamemory: a theoretical framework and new findings, in T.O. Nelson (ed.) *Metacognition, Core Readings*, 117-130.
- Parnas, J. & Sass, A. (2001). Self, Solipsism and Schizophrenic Delusions. *Philosophy, Psychiatry & psychology*, 8, 2-3, 101-120.
- Peacocke, C. (1998). Conscious attitudes and self-knowledge, in C. Wright, B.C. Smith & C. MacDonald, *Knowing our own minds*, Oxford, Clarendon Press, 63-98.
- Perry, J. (2000). *The Problem of the Essential Indexical and Other essays*.Stanford CSLI.

- Proust, J. (2000). "Awareness of Agency : Three Levels of Analysis", In T. Metzinger (ed.), *The Neural Correlates of Consciousness*, Cambridge, MIT Press, 307-324.
- Proust, J. (2001). A plea for mental acts, *Synthese*, 2001, 129, 105-128.
- Proust, J. (2002). A critical review of G.Lynn Stephens & G. Graham's *When self-consciousness breaks*, *Philosophical Psychology*, vol. 15, 4, 2002, 543-550.
- Proust, J. (2002). Can "radical" theories of simulation explain mental concept acquisition ? in J. Dokic, & J. Proust (eds.), *Simulation and knowledge of action*, Amsterdam : John Benjamins, 201-228.
- Proust, J. (2003). Does metacognition necessarily involve metarepresentation ? *Behavior and Brain Sciences*, 26,3 : 352.
- Proust, J. (2003). Thinking of oneself as the same, *Consciousness and cognition*, 12, 4, 495-509.
- Proust, J. (in press). Rationality and metacognition in non-human animals, in S. Hurley & M. Nudds (eds.), *Rational Animals ?*, Oxford, Oxford University Press.
- Rizzolatti, G. & Craighero, L.(2004). The Mirror-Neuron System, *Annual Reviews Neuroscience*, 27:169-92.
- Rochat, Ph. (2003). Five levels of self-awareness as they unfold early in life, *Consciousness and cognition*, 12, 4, 717-731.
- Rosenthal, D. (1993). Thinking that one thinks, in M. Davies & G.W. Humphreys (eds.), *Psychological and Philosophical Essays*, Oxford, Blackwell, 197-223.
- Ruby, P. & Decety J. (2003). Effect of perspective taking during simulation of action a PET investigation of agency. *Nature Neuroscience* : 4,5, 546-550
- Shallice, T. & Burgess, P. (1991). Higher-Order Cognitive Impairments and Frontal Lobe Lesions in Man, in H.S. Levin, H.M. Eisenberg & A.L.
- Shallice, T., *From Neuropsychology to Mental Structure*, Cambridge, Cambridge University Press, 1988.
- Smith, J. D., Shields, W. E., & Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*, 26,3 317- 373.
- Spence, S.A., Brooks, D.J., Hirsch, S.R., Liddle, P.F., Meehan, J. Grasby, P.M. (1997). A PET study of voluntary movement inpatients with schizophrenia experiencing passivity phenomena (delusions of alien control). *Brain*, 120: 1997-2011.

Stephens G.L. & Graham, G. (2000). *When self-consciousness breaks*, Cambridge, Mas. : MIT Press.

Wolpert, D.M., Ghahramani, Z. & Flanagan, J.R. (2001). Perspectives and problems in motor learning, *Trends in Cognitive Sciences*, 5,11: 487-94.

Wolpert, D.M., Ghahramani, Z. & Jordan, M.I. (1995). An internal model for sensorimotor integration, *Science*, 269 :1880-1882.

Wolpert, D.M., Miall, R.C. & Kawato, M. (1998). Internal Models in the cerebellum, *Trends in Cognitive Sciences*,2,9: 338-47.