Thinking of oneself as the same

***Consciousness and Cognition*, 2003, 12, 4, 495-509.**

Joëlle Proust

Institut Jean-Nicod (CNRS, Paris)

I

In which conditions can a person be constituted and recognize herself as herself ?

What is a person, and how can a person come to know that she is a person? Several answers have been explored by philosophers, - having an individual body, and individual brain, having specific introspective access to one's thoughts. They all turned out to be non-starters. A major reason why they do not work, is that they fail to account in a non-circular way for the fact that a person is inherently both a stable and a changing entity; an entity, furthermore, who knows herself as herself. If the essence of a person is to be an historical object, a " continuant ", it follows that the only ability through which a person can be revealed to herself is *memory*. Locke gives us the following indication:

As far as any intelligent being can repeat the idea of any past action with the same consciousness it had of it at first, and with the same consciousness it has of any present action, so far it is the same *personal self. (Essay*, II, XXVII, 10)

Now this identity between a consciousness that " repeats " a past action and the consciousness that accomplished it involves an interesting semantic property. To reach knowledge of oneself as oneself, more than a simple factual identification to an " I then " with an " I now " is required*. What is further needed is that the " I " is recognized as the same by himself across these two cases.* Let us take for example a memory in which I recall that I visited the Versailles castle. It is not sufficient that the I in " I recall " and the I in "I visited the Versailles castle " happen to refer to the same person. I must in addition *know* that both tokens of " I " refer to one and the same person, me. Contrast this with the use of the third-person pronoun in the

following sentence : “ John thinks about his father ; he remembers the day when he died”. The first “ he ” refers to John, the second refers to his father. There is no co-reference in the sentence.

One might think that in the case of “ I ”, two tokens must necessarily co-refer when they are thought by the same thinker. That it is not necessarily the case, can be seen if you take, for example, two messages of your computer : “ I need help ”, “ I found three answers to your query ”. These two instances of “ I ” clearly do not need to include conscious co-reference : the message is conveyed to you even though your computer has no specific self-representation constituting the unitary framework for the two usages. What applies to the computer may also apply to usages of the first-person pronoun in language-instructed apes, in young children or in patients with neurological disorders. Hector-Neri Castaneda called[1] “ quasi-indexical usage ”, noted “I*”, the application of the first-person pronoun when there is a recognition of the co-reference between the two tokens of “ I ” in such contexts as reported above (“ oblique contexts ”). In I* cases, the subject who forms the belief and the subject to whom a property is attributed (in the belief) are recognized as identical. Without such a capacity to refer through a designator that relates reflexively two contexts with an I-tag, as in “ I believe that I did it, that I saw it, ” etc., one might acquire types of information that in fact (or implicitly) are about myself, but fail to know explicitly that it is about myself that they are.

It is thus clear that instantaneous types of self-awareness as can be offered in perceiving or acting cannot suffice to *give us access to* a self as identical to him/herself over time. As an invariant, a self cannot be reached in a single snapshot. This epistemological claim is related to a corresponding metaphysical claim: a person cannot *exist* aside from a historical process, such that a sequence of cognitive states allow this or that personal identity to emerge. To be a person, one needs minimally to be conscious of two different things *and* to bring these two contents together in the same present conscious experience[2]. This kind of co-consciousness involves more than lining up various attributions to myself (for example, "I remember that I visited Versailles Castle; I am now looking at the picture I then took"). It requires a capacity

1 Castañeda, (1994).

[2] This observation does not imply that properties made available in perceptual experience (whether proprioceptive, visual or auditory) and in the experience of acting cannot be included in the consciousness one has of being identical to oneself. More on this later.

to recognize the identity between the " I " as a conscious subject and the " I " as the topic of a report, a memory, etc.

If one now decides to offer an account of persons in terms of individual memory, two things have to be done. One consists in examining whether selves are *bona fide* entities, and, if they are, in showing what they consist in. The other is to explain how one gets access to the self one is, or is supposed to be - without involving any circular reference to a previously introduced self. It is worth briefly recalling Locke's own claim that conscious memory of prior perception and action constitutes personal identity, and show why it fails to provide the kind of non-circular account we are after. In Locke's " simple memory theory " (as we will call it), being a person simply depends on the continuity of memories that an individual can bring to her consciousness. Even if we don't recall all the facts of our past lives, memories do overlap, which delineates the extension of a continuing person.

II Problems of the simple memory theory

Locke's definition of a person raises various problems– some of which have been solved. We will have to summarize them briefly in order to capitalize on the results of these classical discussions. A human being, Thomas Reid observes, generally cannot remember all the events of his life. The old general remembers having been a brave officer, and as a brave officer he could remember having been whipped for stealing apples when a child. But the old general does not remember the child's being whipped. Reid concludes that, according to Locke, the old general both is and is not the same person as the whipped child.

A simple memory theory is also relying on a quite obvious kind of circularity[3]. In requesting that the appropriate way in which S's memory was caused should be one in which S himself observed or brought about an event, one insists implicitly that the person who remembers is the *very same person* that witnessed the event or acted in it. How might a self ever be constituted, if one already needs to reidentify *oneself* through memory to get self-cognition under way?

3 See on this question Shoemaker (1970), Parfit (1984), and my Proust (1996).

Sydney Shoemaker offered an interesting, if controversial, solution to cope with these two difficulties. He defines the psychological state of "having an *apparent* memory from the inside" as what one has when the concrete memory of an event jumps to mind, in contrast to memories that do not involve any direct participation. For example, you may remember 'from the inside" witnessing the coronation of Queen Elisabeth the second, as it was an experience you may have had. Whereas you cannot remember from the inside the coronation of Carlus Magnus. Thus characterized, (that is, by definition), being in this state does not presuppose necessarily that one is the person who actually had the experience. The general strategy is to define true memory on the basis of apparent memory (the subjective impression of having had an experience), and to build the notion of a person through a succession of overlapping apparent memories. In this view, personal identity cannot consist in remembering all the events of one's life, but in an ordered relation between direct rememberings, such that each one is connected to the few previous ones. In Parfit's version of this improved definition, there is a person when there are "overlapping chains of direct memory connections".[4]

Another problem however is raised by the simple theory, as well as by the versions just sketched. It is connected to one of the consequences of the quasi-indexical nature, reflexive meaning of the " I* ", namely the unicity of the I*-thinker. In order to constitute personal identity through overlapping memories, we need to secure the quasi-indexical property of the two tokens of " I ": the one who remembers and the one who initially acted or perceived. But even though continuity of memory is realized, there is no conceptual guarantee that there is *only one man* who fulfills the memory condition imposed for being the same person. Leibniz seems to have been the first to underline this difficulty.[5] He reasons in the following way. Let us suppose that there is a twin earth, crowded with people exactly similar to the inhabitants of this earth. Each of a pair of twins will have all the physical and mental properties of the other, *including the same memories*. Are they two persons or one? Clearly, Leibniz observes, an omniscient being such as God would see the spatial relation between the two planets, and discriminate them. But if we stick to a mental property narrowly

4 Parfit, (1984), p. 205.
5 Leibniz (New Essay), II, 27, 23.

conceived such as memory, that is, the consciousness of having seen or acted, there will be no way of justifying the intuition that a person is unique.

This so-called " reduplication " argument can be generalized to all the definitions of personal identity that rely on a " Cartesian " psychological property, that is, a property of the individual's mental states individuated in a purely functional way (independently of the context of her thoughts and memories that – in externalist views of meaning -, contribute to the very content of what she thinks). Inserting copies of the same set of memories in previously " washed " brains would result in the same kind of reduplication as the twin earth story[6]. The problem is not, of course, that such a circumstance is around the corner, but that there is no conceptual answer to the Leibnizian question: how many persons are there? One? Two? More? And if the copying is imperfect, can one say that two of the clones in the same set are " approximately " the same person?

The simple memory theory has more recently revived in narrative views of the self, defended either in the context of an artefactual theory of the self (Dennett, 1991), or as a substantial, hermeneutic theory of human persons (Ricoeur, 1990, Gallagher, 2000). Each of us is supposed to reconstruct, or have access to the self he/she is by unpacking retrospectively his/her particular memory sequence. This version belongs however to the class of simple memory theories and therefore, falls prey to the reduplication argument. Moreover, the kind of narrative that is selected suggests at best an idealized view of oneself, reduced to the requirements of story telling (avoid redundancy, only select salient facts, produce retrospective continuity, rely on the benefits of hindsight). Thus, the descriptive condition on memory overlap combines freely, on this view, with a normative dimension of which memories are worth contributing to one's self-narrative; this normative dimension may create a difficulty if one defends a realist view on the self. For it is difficult to dissociate it from the genre of story telling; the self appears clearly as a fictional entity reconstructible in variable ways according to present values. These two observations suggest that the narrative view on the self is more consonant with a reductionist metaphysics - one in which there is no self to know, and in which the self just is the story that an individual human being is telling on her prior life at a given time.

---

[6] This observation led Velleman (1996) to distinguish selfhood, defined perspectivally (as remembering and anticipating experiences first personally) with the identity of a person. See in particular p. 66 and and p. 75, fn 53.

One interesting feature of the narrative view, however, is that it highlights a possible function of self-focused memory. Rather than being a detached contemplation, memory appears as an active, retrospective evaluation of former actions and perceived events, with en eye on future actions and dispositions to act.[7] This feature may not disqualify memory from contributing to individuating selves; but the type of memory required by a realist approach is supposed to shape a working self, not a decorative entity. (Another way to convey this point is to say that a realist on the self is interested in the ontology of self, not in self-ideology). To prevent arbitrary focusing on specific doings or character traits as in a self-narrative, the kind of transformation occurring in self-related memory must express directly the normative-directive function of memory in connection to intentions and plans; it must be an internal, not an accidental feature of the memory process. Finally, the type of memory involved must also be such as to avoid reduplication: it must be not only a mental process that activates psychological states in a subject (I remember doing or seeing this and that), but it should secure the numerical identity of the person. What kind of mental process might fill these requirements?

We just saw that in order to obtain the strong reflexivity of I* that is needed for self-reidentification, memory must participate actively in transforming the organism through the very process of remembering. The form of memory that considers the past in the light of an individual's present worries, and that aims at actively transforming the individual's mind in part through that memory, is the memory involved in *mental action* – a form of metacognition. The claim defended here will accordingly be that mental action alone can deliver the required temporal and dynamical properties that tie the relevant remembered episodes together. Constituting a person presupposes the capacity to act mentally, that is, to consciously monitor and control one's own mental states on the basis of one's past experiences and of one's projects. For such a conscious monitoring of mental actions to occur, a specific capacity must develop over a lifetime. Monitoring one's mental actions consists in rationally revising – and adequately refocusing - one's prior

[7] This dimension did not escape Locke's attention : " Wherever a man finds what he calls *himself,* there, I think, another may say is *the same person*. It is a forensic term, appropriating actions and their merit, and so belongs only to intelligent agents, capable of a law, and happiness and misery ". Locke, *Essay*, II, XXVII, 26, t. I, p. 291. This observation suggests that the self is not only a matter of private enjoyment. From a sociological viewpoint, one would claim that its function is to distinguish and stabilize statuses and roles in a social body, as well as to apply the gratifications and the sanctions inherent to social control. For lack of space, we will not examine further this aspect of selves in the present article.

dispositions to act, to plan, to remember, to reason, to care, and to reach emotional stability. Memory plays a central role in this form of normative metacognition; although philosophers who have studied memory may not have realized this, "memory" is involved in most types of control. Thus, using a philosophical jargon, to *be* a self presupposes a capacity of self-affection. Self-memory is the dynamical ability of modifying one's states deliberately to reach new states that are seen as more desirable. Our claim will be that *an individual's way of gaining both a self and an access to it should be constituted not by the process of recalling alone, but by being conscious of being affected, or transformed, through that very process*. This new hypothesis will be called " the revised memory theory ".

III

The revised memory theory of personal identity

What is a mental action? To understand what it consists in, it is useful to compare it with a physical action. Let us consider an example. When you are training yourself in a sport, you put your body in a condition to fulfill new functions that you find desirable; in tennis, for example, you aim at learning how to execute certain kinds of gestures, like a half-volley or a topspin forehand; you follow all the steps instrumental to reach these goals, i.e. by observing others performing the gestures correctly, by modifying your own bodily attitudes and by discriminating various relevant new properties in the objects involved (the ball, the racket, etc.).

Mental action is very similar to physical action; but instead of modifying physical objects in space, what it aims at modifying are mental states in the agent. In spite of all the efforts aimed at bending spoons, it is clear that there is only one thing that can be transformed through mental action, that is the very mind of the agent who acts mentally. Nor is mental action something difficult or exceptional, requiring a specific training or mediumnic capacities. It only requires using one's past experience to monitor actively one's informational or emotional content: modify one's knowledge (to learn), one's desires (to become cultivated, to become expert in a field[8]), one's

[8] Harry Frankfurt (1988) develops the view that second-order volitions are fundamental for a person to come to existence. His view differs from the present one in so far as self-reidentification is based on a process of "identifying with first-order volitions", while here more general revisional processes are taken to provide the

emotions (to become harder-hearted or to mellow). Mental actions may also be required to monitor one's attention, one's motivations (I first finish the book before I give my phone call), one's addictions (I will smoke only one cigarette before lunch). Therefore, mental actions play a fundamental role in shaping one's life. They make possible the capacity to govern oneself, to reorient the course of one's thoughts, one's desires, one's learning; they allow for the adjustment of motivation and effort, for persistence or change in love, seduction and disgust, for the choice of a field of activity and for the scope of one's responsibility. All these actions can be redescribed as self-monitoring for the benefit of self-control; from an initial mental state, and a particular set of dispositions, they are needed to actively acquire new mental states (new contents or new attitudes to old contents) and dispositions.

Constituting a self

In short, self-affection refers to the general ability of taking new mental states and properties as goals of action, and of pursuing these goals. Given such an ability, the sense of being oneself, a person identical over time, with the strong reflexivity on which the notion of a person depends, *consists in the ability to consciously affect oneself: in the memory of having affected oneself, joint to the consciousness of being able to affect oneself again.*

In the context of a discussion of Frankfurt's view on the self, J. David Velleman (2002, 111) observes that the reflexivity of the control exerted over one's own behavior does not offer access to a single entity. The present paper argues however that there is a level of control that needs to put all the various reflexive mental states (perceptions, intentions, thinking episodes) in harmony for a consistent interaction with other agents and with nature to be at all possible. The emergence of the self in phylogeny might reflect the extension of human memory, compared to other primates; verbal communication allows commitments to be taken, social roles to be attributed, as well as sophisticated plans - social or technological). The mind of an individual participating in this kind of communication and action needs to adjust flexibly to new tasks and environments (it must change itself without losing track of its

---

functional condition for self-reidentification. Furthermore, the present view rejects the claim that a person has an individual essence based on the motives she identified with. See Velleman (2002) for an interesting discussion on this point.

own previous states). The self is the dynamic function that results from this set of selective pressures. While self-conceptions may considerably vary from one society to another, the structure that is here described under the term of "self" is a universal feature of our species.

It is an *a priori* necessity that a mental agent permanently monitors and changes her own knowledge state, her own emotions, or her own present conduct. In other words, a mental agent "cares about" her mental life (to borrow another Frankfurt's expression), and knows - at least in a practical way - how mental properties (the amount and quality of knowledge reached, or attention focused, of motivation gathered) affect her dealings with the external world. In the present perspective, the type of mental structure on which a self supervenes is gained in the exercise of the disposition to act mentally over a life sequence. The overlapping memory episodes involved in this process provide the kind of continuous link needed for reidentification, just as in simple memory theories. Contrary to these, however, only mental agents may qualify for selfhood; agents able to form memories of their previous actions and observations, but not to act mentally – those restricted to physical actions - do not develop into persons.[9] Individuals of this kind would be unable to resist their impulsions, as is the case for Harry Frankfurt's *wantons,* for lack of any control on what they think and do.[10] The reason why these individuals do not qualify for selfhood is not that they cannot access their own bodily condition and care for it, not that they cannot remember how their body, or prior physical condition was (these types of knowledge are indeed present in non-human primates, a good illustration of wantons). It is that they fail to monitor their long-term dispositions, revise their beliefs or plans. If they neither understand, nor care, for the consequences that courses of action have on their control capacity, they cannot reorganize their preferences in the light of overall constraints. "Self" thus designates an endogenous individual structure of the will based on a form of metacognitive memory.

Note moreover that reflexivity and, consequently, numerical identity are *intrinsic* to the permanent revisional process in which acting mentally consists. This is crucial to prevent the reduplication problem. Even if, at a given moment, an individual's thought was copied into another's mind, each clone would re-individuate herself through the practical reflexivity that governs her mental actions; as soon as each

[9] See Proust (2000) for a thought experiment to this effect.

agent has revised her beliefs to act in her own sphere, with her own body, she has become a person with the double backward/forward dimensions of reidentification that are open to her.

What about the normative dimension of selves? Clearly, the capacity to remember how one acted, joint to the capacity to change one's plans, open up opportunities for commitment. The self is constituted by the normative commitment that an agent (not a self, yet, let's call the relevant instance of agency: a mind) having information on her states has to *revis*e her dispositions - if incoherence or means-end inadequacy arises, or in the case in which there is a conflict in her habits and her preferences,) and to offer (to herself or to others) a *justification* of what she did in terms of the content of her attitudes in relation to a goal. *"Justification"* should be understood here in a minimal way: the agent just aims at behaving rationally, in the sense that she does not want to lose her physical and mental resources on goals that are detrimental, meaningless to her, or impossible to reach. In others words, an agent tends to act on the basis of her preferences. An important thing to observe at this point, is that most (physical, ordinary) actions presuppose a capacity for mental action; they require planning and deliberation, emotional control, directed learning and other forms of memory control; they are effected through a mental simulation that may itself be rehearsed and modified.

We are now in a position to respond to Velleman: How do we get unicity in the mental organization if not from the coincidence between the mind that revises and the mind that is being revised? Selves are results of metacognitive processes in which minds reorganize themselves to be more attuned to their physical and social worlds. The revised memory theory of selfhood therefore suggests that to *be* a self, an agent must

a) Be capable of metacognition, which includes (inter alia): forming dynamic mental goals, i.e. appreciating, adjusting and revising prior preferences about mental goals.

b) Form overlapping memories of previous revision episodes.

c) Reorient one's own mental actions on the basis of a and b; revisions are used in the course of planning overlapping future courses of mental actions.

---

[10] See Frankfurt, 1988.

Accessing the self

Now let us turn to the second question we had to answer: how can a mental agent get access to the self that emerges from her dispositions to act mentally? In order to give a clear answer to the question, we need to develop the distinction between control and monitoring – two dimensions that have to be present in an organism capable of autonomous action. Any control process, however complex, is composed of a two-phased cycle; in the efferent phase, a command based on an internal simulation of what is to be achieved (providing a form of expectation of what the environment is like) is sent to the active organs (muscles, or - in case of a mental action - internal thought processes); the second phase gathers information and possibly replaces the anticipated with the observed feedback for the sake of further control purposes[11]. If such is the functional division of any control structure, the self *exists* in virtue of the whole structure, and the question whether it belongs rather to the control level, where the norms are constructed and used in prediction, or to the observed feedback level, where the actual evidence is sampled for further revision of former plans, does not need to be reflected in two independent objective dimensions *inside* the self. Just as in the case of normal action, you don't need to distinguish how the action was programmed from how it eventually went (the two courses are more or less coextensive; what matters is that correction *can* be executed in case unexpected things happen).

Now the question of how an individual *gets access* to herself presents us with the possibility of two choices – a possibility that is interesting for psychological, sociological and for moral reasons. As we saw above, different cultures have various ways of conceiving what a self is. Given the two-phase (control+monitoring) cycle of any autonomous system of this kind, two emerging structures might offer access to selfhood. There is the level of what is ideally aimed at (the control structure); and there is the level of what is observed (the monitoring evidence). These two levels might play a distinctive role as far as access is concerned; they seem to match

[11] In the particular case of chosing courses of actions, the cycle control –monitoring is temporally extended over sequences of varying duration (think of when you decided to be, say, a philosopher, and when you started to get internal feedback on being able to reach this goal).

respectively the notions of an ideal vs. an actual self: what the individual sees herself as striving for becoming, vs. what the individual sees herself as in fact being.

Obviously people may misrepresent who they are. Self-conception is not constrained by perceptual or introspective mechanisms. Although metacognition offers both implicit and explicit forms of access to previous revision episodes, in particular the most salient and long-term ones, an individual may be delusional, or simply confused, about who she is. Nothing prevents an individual (who may even be a "wanton") to take herself as aiming at things that she actually does *not* incorporate into her control system.[12]

One might at this point speculate that each particular culture frames selves in specific external signs that somewhat co vary with the metacognitive ability that our analysis has pointed out as being the basis of selves[13]. We might further speculate that each human individual can come to understand from her own practice of mental action how her mind develops into a self, or unfortunately also, can dissolve away from a self, when the conditions are present. As we will indicate below, this capacity is certainly fueled by using words - in a public language, like “I”, “you”, etc., or proper names - that express the normative and descriptive aspects linked to selfhood. Use of these words is naturally part of an overall social structure that may, or not, encourage individual beings to take responsibility for their own choices and stimulate their own autonomy in revisional practices.

Now some readers may at this point worry that the present suggestion does not escape a form of circularity. Here is how the objection might go. One of the most common ways of acting mentally consists in *revising* one’s beliefs, desires, and one’s commitments, as a consequence of changing circumstances and incoming factual knowledge. It is in this activity that a person is supposed to be built: to know who one is, it to know, in a practical and concrete way, the global target and the stake of the revisions and adjustments that have been or are being currently performed in the

[12] Many individuals might thus capitalize on their expected, or simply imagined, mental agency rather than on their actual evidence for being mental agents capable of revision. The story-telling evoked earlier might induce them in believing, for example, that they are better planners of their own lives than they actually are. Others might collect comparative or social evidence (diplomas, external recognition, friendly reports) for ascertaining which kind of self they have. All these individuals would thus lack knowledge of who they are, because the proper form of access is located in the reflective sphere that controls preferences and plan revision rather than in public achievements.

[13] See Goffman, (1959).

domain of mental action. Now one might express here the worry that such a mental kind of activity does not constitute selfhood or personal identity, but rather *relies on* it. For is not the individual mental agent already implicitly taken to be/have a self? Is not this latter condition presupposed for re-identification to occur through the revision process? When an agent takes a revisional commitment, and engages her future on the basis of an evaluation of what she can and should do, given her present global judgment, is not her own self causing the utterance? So how could it make sense to extract, so to speak, selfhood from mental agency?

To address this objection, one has to contrast the objector's notion that an action is caused by a person, with the naturalistic analysis of action. On the latter view, the agent is not supposed to have a causal role in her actions: her intentional states, or the external properties of the environment do. It is natural to say that an agent is responsible for her actions; but at present our theoretical interest is of another, more fine-grained sort: we have to provide the definition of a self. And the only way of doing so is to rely on a subset of her intentional states that do have a causal role in her actions and that warrant the reflexivity of I*. Why cannot we identify the self with *all* the agent's intentional states? First, because they are an unstable, moving, heterogeneous crowd, -- all but distinctive of this person: look how widely shared are the likes, the dislikes, the emotions, the beliefs, etc. of each one of us! But also because intentional states can in principle, as we saw, be copied, and made a priori to characterize several distinct individuals, which leaves us with an undetermined concept of self. Our definition is thus not circular. No self is presupposed before the reflexive intervention of revision gets into play. The self results from this reflexive process, and will stop developing, or decay, if the reflexive process is interrupted, or is reduced.

To be a person, in this analysis, can thus be reduced to the exercise of a disposition to act mentally. Such a reduction does not aim at " doing away " with persons, however. Persons may not be fictions - not *only* a matter of " self-presentation ". For when somebody pretends to be someone he is not, he is still expressing his actual capacity at revising and planning. Nor are they substances, something that remains to be known, observed, made explicit. A person is a system

of dispositions, socially encouraged and trained, designed to revise beliefs, desires, intentions, and thereby become the actor/goal/target/ of one's own life.

IV - Pathologies of the self

Recent work on personal identity has attracted philosophers' attention to the schizophrenic delusions; deluded patients indeed seem to change their minds not only on their own personalities, occupations and capacities, but also on the very extension of their selves. Some are intimately convinced that they are deprived of a self and do not know how this word might refer at all; the word seems to them to provide an artificial unity to a bunch of multiple and disconnected mental experiences. Other patients, in contrast, feel included in a wider personal entity that encompasses not only their own minds, but also others' as well.[14] The sense of a lost or of a transformed self is associated with an impression of "extraneity" in thought and/or in action: these patients have the feeling that their actions are controlled by others, or that their thoughts are inserted in their minds from without. Such cases seem to suggest that, contrary to traditional claims, one can be wrong about who one is.[15] There is no "immunity to error through misidentification".[16]

---

[14] A deluded patient for example claims "I am you (pointing to John) and you (pointing to Peter)"; an other patient describes his inner experience in the following terms "Thoughts have been put in my head that do not belong to me. They tell me to dress. They check the bath water by doing this gesture".

15 See in particular Campbell (1999, 2002), Proust (2000b), Gallagher (2000), Stephens and Graham (2002).

[16] Such immunity was traditionally thought to apply to the usages of self-referring terms such as "I" ; it consists in the impossibility of being mistaken about the person who employs the word "I". What is meant by that, is that there is an essential asymmetry between the usage of "I" and of other singular personal pronouns, such as "you" or "he/she/it". I can for example use the word "he" mistakenly, either because the person designated is in fact a woman, or because what I point to is actually the shadow of a tree. I can also say mistakenly "you" to something with no self.[16] When someone says "I", however, the reference seems to be immediately secured and accessible, i.e. without needing any mediating property to know who « I » could possibly be; besides, it does not seem open to a thinker to be wrong about whom he means to designate in this way; for this reason, philosophers have concluded that self-attribution in thought has a special epistemological status : you may certainly be wrong on many of your own characteristics, personality traits, etc., but never on whom you pick up with the pronoun "I". This again suggests that in order to refer to yourself, you don't need to identify a person among others, i.e. yourself rather that this or that other possible "ego". For if you had to use some identifying feature to know who you are in the other kinds of personal attribution, you could in principle - sometimes at least - be wrong about whom you pick up. As it seems that you can never be wrong about that, it follows that you have to refer to yourself in a way unmediated by identification.

We now have to turn to the empirical evidence from neuroscience that might contribute to explain these symptoms: is the present approach compatible with it? How does our effort at clarifying the conceptual analysis of self-identity fare with the scientific analysis of the mechanisms involved in perturbations of the self? An influential view in the neurophysiology of schizophrenia is that the capacity to self-attribute a thought, an intention or an action, is dependent upon the control of agency[17]. There are at least three different ways of articulating this idea.

a) Chris Frith's most recent view is that a breakdown in the mechanism of efferent copy and comparator results in the breakdown in the sense of agency. Schizophrenic patients seem to be unable to monitor their own motor instructions (p.614). Many studies[18] have shown that they rely on visual feedback rather than on the efference copy of the motor commands to predict the success of their physical actions; in my terms: they apply a form of control named "error-control ", instead of applying a "cause-control")[19].

b) An earlier view, also defended by Frith[20], was that the capacity to act in agreement with one's intentions - in particular, when the intentions are "endogenously generated", rather than stimulus-driven (triggered by a routine feature of the context), requires a capacity to attribute these intentions to oneself. On this view, to be able to act properly, you need to use a theory of mind, and metarepresent your own states, to understand that you are the agent of your actions and the thinker of your thoughts.

c) Marc Jeannerod's view is that a common covert simulatory process that is activated both when you see someone act or when you act yourself, generating shared representations of actions[21]. The process through which you attribute an action to the self or to another agent is explained not at the level of the action-monitoring system, as Frith does, but at the level of the simulation mechanisms involved in acting and in observing actions: an inability to simulate correctly the covert operations involved in attributing

---

[17] In schizophrenic patients, the sense of willful activity seems to be perturbed simultaneously at three main levels i) in the pursuit of reward that structures behavior and goal hierarchy (basal ganglia), ii) in the imagination and in the selection of novel actions (dorso-lateral prefrontal cortex, left side), and finally iii) in the attribution to self of the effects of an action (right inferior parietal lobule and superior colliculus). These three dimensions are closely involved in the revision process that have been described above.

[18] See for example Frith & Done, 1989, Mlakar et al., 1994.

[19] On this distinction, cf. Conant & Ashby 1970.

[20] See for example Frith, 1992.

[21] See Jeannerod, 1999, Jeannerod and Pacherie, submitted.

actions either to self or to another agent, explains why the patient has an impression of external control.

What is worth observing, first, is that these three views on self-attribution are explicitly or not "control theories" of the self. In the first view, self-attribution of action is mainly secured by the forward "motor" model through which the motor and distal consequences of the action are predicted. In the second, control is operated through a propositional metarepresentation of the first order intention to act. In the third, control is operated off-line, in covert simulations, and what is perturbed lies in monitoring the covert reafferences of this postulated additional control system.

It is to be noted, second, that the three views above do not try to understand how a self-representation is *generated*, but with how an action is *self-attributed.* This latter task may involve, at best, access to a previously acquired self-representation; but the theories above do not aim at establishing whether (and how) a permanent self can be accessed from one self-attribution to another. The same thing applies to most discussions of the relevance of pathology to solve the puzzle of immunity to error to misidentification of I-thoughts. The whole debate had the merit to stress the difference between a sense of subjectivity, (or of ownership), through which an individual has the subjective experience of thinking, perceiving or acting, on the one hand, and the sense of agency, (or of copyright), in which the subject feels that she is the author of her act or the thinker of her thought[22]. But the way the distinction is used, (and implicitly restricted to humans) in the debate presupposes that we already have identified the *stable* foundation on which a subject can establish the sense of being the *same* self - a basis that is crucial, as we saw, not only for the possibility of *reidentification,* but also for the *unity* of a self at any given time[23].

The concept of self-attribution is thus ambiguous, between a "human" self and a lower-level « motor control » self. It may refer, on the one hand, to the sense that an occurrent action or thought is under one's control - a sense of agency that we share with other animals (primates, mammals, birds and fish); or it may mean, on the other hand, that the agent reflects on her actions as an expression of her long-term beliefs,

[22] One of the questions that we can clarify on the basis of the present approach, is how the possession of a self, joint to the sense of being oneself, interacts with the sense of agency and with the sense of ownership. This is a complex question that we explore elsewhere.

[23] See Peacocke, 1999 Campbell, 1999 and Proust, 2000b, for an elaboration of this point.

and gets control on her motivations, in a more unified and "interwoven" way.[24] Several authors have analyzed this interwovenness as the recognition that our occurrent thoughts are causally determined by our long-standing propositional states[25]. But as we saw, this will not secure the unicity of the thinker. The thread of a self does not consist in belief possession, (subject to reduplication), but rather in self-affection, that is in the capacity for a single occurrent thought to deliberately transform not only other states, but also mental dispositions.

Given the ambiguity explained above, most authors are in fact dealing with a form of self that has nothing to do with a reidentifiable self; they are interested in attributions of agency of the type "I willfully broke the vase" versus "I was pushed and broke the vase". What our former conceptual analysis suggests however is 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, is located at a level distinct both from the control loops of unreflective action, perception, and memory, and from the level of simulating and planning actions. There must exist a *third* level of control, at which a subject is able to simulate and monitor not her elementary perceptions and actions, not her plans, but the courses of revision needed for the viability of the agent among other agents, in a context extended over time. The subject needs to form dynamic models of her own mental dispositions, to keep track of her previous revisions and critically examine how reafferences match what was expected. This allows her to plan her life at a deeper level than just the instrumental level engaged in ordinary agency. It also allows her to simulate observed agents engaged in individual, competitive or cooperative tasks, with possibly conflicting intentions or selfish goals.

If our view is correct, there must be a semi-hierarchy of control levels; each is established through a "command and predict" cycle that is functionally associated with other levels of the hierarchy. The term of a semi-hierarchy refers to the fact that the various control loops can work in part independently: you can represent yourself doing something without doing it actually, and you can also act without thinking about the way you do it, or whether your doing it conforms to your long-term goals and

[24] Campbell, 1999, 621.
[25] See, Campbell, *Ibid*., 620., Armstrong,1968, Stephens and Graham, 2000.

values.[26] Therefore a represented self may be a motivation for acting or not in specific ways, but it can also be inactive, or perturbed, without altering ordinary perception and action. Reciprocally, the kind of metacognition relevant for a self is not engaged in every single ordinary action. There may be however specific changes at lower levels that should drastically affect self-recognition.

In a simplified view of the control functions engaged in an entity capable of metacognition, three levels have to be distinguished. Level 1 controls and monitors sensory processing as occurring in perception and in motor activity. Level 2 simulates and monitors agency: various courses of action must be chosen at every single moment; progress towards distal goals has to be monitored until completion. These kinds of operations presuppose some form of hierarchical dependency between level 1 and level 2, although automatic attentional capture must be present to interrupt level 2 control when needed. Level 3 simulates and monitors agentic capacities in the light of long-term values and plans. Again, level 3 operation presupposes that level 2 operations can be relied upon as instantiating level 3 control models.

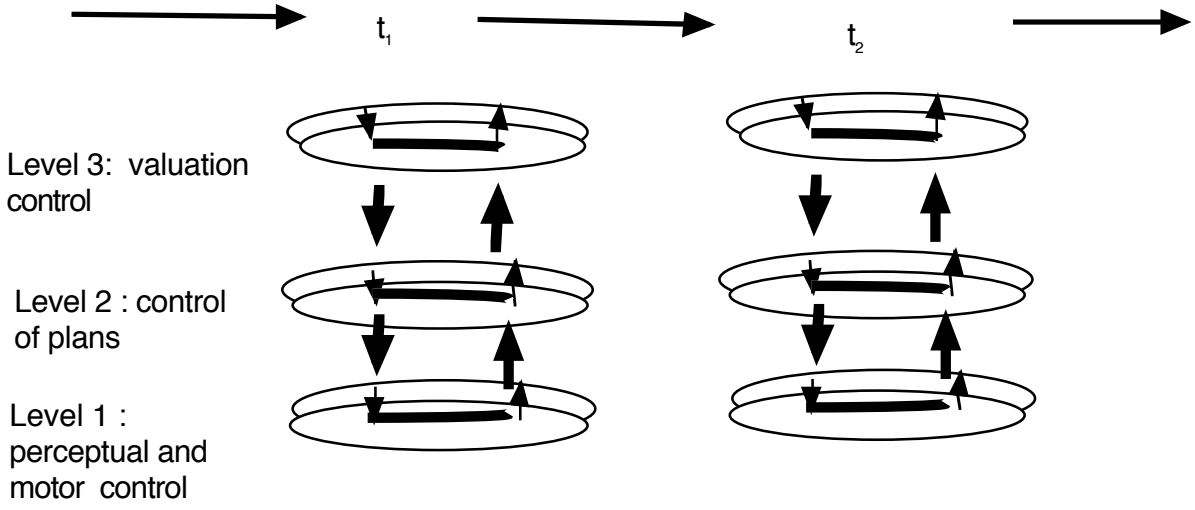

Fig 1: Control levels: a semi-hierarchy.

[26] For example, if you plan to be a pilot, you need to bring yourself to act as a pilot, to perceive as a pilot, etc. Reciprocally, you can realistically plan to become a pilot only if your eyesight is correct etc. The important point is that you do not need to permanently represent yourself as "a pilot" to pilot.

Let us observe that, in such a control theory, representations of possible alternative models of a dynamic evolution can be formed on the basis of endogenous as well as exogenous stimuli. What Jeannerod calls « covert simulation » belongs to each control level insofar as each requires a feed forward model of the developing situation. There are however various ways of covertly simulating a process, according to whether it is a motor, decisional, or evaluative process. Simulation thus has to be referred to a specific functional domain, according to whether an action is to be predicted in its motor (level 1), instrumental (level 2) or social/evaluative consequences (level 3).

Let us see how control theory allows to better understand the findings reported above. If it is correct to claim that, in a schizophrenic patient, level 2 control is disturbed, (agency control, in its conscious monitoring dimension), whereas Level 1 control is untouched, the subject recognizes the reafferences in their subjective component, but without the sense of expecting them as a consequence of her own willful agency. There is therefore a frontal clash between level 1 intuitions of mineness and level 2 intuitions of unwillingness. Level 3 is called upon to provide a dynamic model of the case. Conflict is solved at level 3 in a more or less automatic way: the feeling of being compelled to act provides subjective reafferences for the immediately higher level of control; the subject *senses* her own self being dissolving, parasited. On the reciprocal cases in which a patient attributes to herself control of others' actions (at level 2), her self is felt as amplified and extended to other agents (at level 3). In both cases, deluded patients experience an alteration in the self-other division – either because the self includes other beings (sensed as now being under potential self control), or because the self has become part of other beings, or agents (sensed as taking control of agency). These two forms of self-alteration are generally coexisting with a preserved sense of individual history and memory, as can be predicted in the present hypothesis.

Conclusion

The philosophical problem of personal identity consists in offering a way of defining a self that allows understanding how an individual can be – and represent herself as - the same self although her mental and bodily dispositions vary

considerably, as well as the environment in which she is leading her life. We suggested that this property of “ipseity” (a form of identity compatible with change over time in certain properties) could only be captured in a memory process specializing in dynamic belief/desire and value revision. This capacity belongs to metacognition. Our goal here was first to show on a conceptual basis how self-affection constitutes the only way of constituting a self, and of reidentifying oneself in the strong reflexive sense required. We further sketched how this conceptual structure is realized in a semi-hierarchical control system; its control and monitoring dimensions account for the normative and descriptive components in self-representation. Finally, we briefly indicated how this account allows clarifying the discussion of schizophrenic symptoms relating to self.

References

Armstrong, D. (1968), *A Materialist Theory of the Mind,* London, Routledge and Kegan Paul.

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 and Psychology*, 9, 1, 35-39**.**

Castaneda, H.-N. (1994), "On the Phenomeno-Logic of the I", dans Q.Cassam, (dir.), *Self-Knowledge*, Oxford, Oxford University Press, 160-166.

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

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 schizophrenic patients, *Cognition.* Vol. 65, pp. 71- 86.

Dokic, J. & Proust, J. (eds.) (2002). *Simulation and knowledge of action,* Amsterdam : John Benjamins.

Frankfurt, H.(1988) *The importance of what we care about*, Cambridge, Cambridge University Press.

Frith C.D. (1992), *The cognitive Neuropsychology of Schizophrenia*, Hillsdale, Lawrence Erlbaum Associates.

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.

Frith, C.D. & Done, D.J. (1989), Experiences of alien control in schizophrenia reflect a disorder in the central monitoring of action, *Psychological Medicine*,19, 359-363.b
Gallagher, S. (2000). Self reference and schizophrenia, in D. Zahavi (ed.), *Exploring the self*, Amsterdam, John Benjamins, 203-239.
Goffman E. (1959), *The presentation of self in everyday life*, London : Penguin Press.
Hume, D., (1888). *Treatise on Human Nature*, L.A. Selby-Bigge (Ed.). Oxford : Oxford University Press.
Jeannerod, M. (1999), To act or not to act, perspectives on the representation of actions. *Quarterly Journal of Experimental Psychology*.52A :1-29.
Jeannerod, M. et Pacherie, E. (submitted), Agency, simulation and self-identification.
Leibniz, G.W. ([1705]-1997) *New Essay on Human Understanding*, transl. & edited by P. Remnant & J. Bennett, Cambridge, Cambridge University Press.
Locke, J. ([1695]-1971) *An Essay concerning Human Understanding*, London, Dent.
Mlakar, J., Jensterle, J. Frith, C.D. (1994), Central monitoring deficiency and schizophrenic symptoms, *Psychological Medicine*, 24, 557-564.
Parfit, D. (1984), *Reasons and Persons,* Oxford, Clarendon Press.
Peacocke, C. (1999), *Being Known,* Oxford, Clarendon Press.
Perry, J. (1975), *Personal Identity*, Berkeley, University of California Press.
Proust, J. (1996), Identité personnelle et pathologie de l'action, in I. Joseph & J. Proust (eds.), *La Folie dans la Place, Pathologies de l'interaction, Raisons Pratiques,* 7, pp. 155-176.
Proust, J. (2000a), "Awareness of Agency : Three Levels of Analysis", In T. Metzinger (ed.), *The Neural Correlates of Consciousness,* Cambridge, MIT Press, 307-324.
Proust, J. (2000b), "Les conditions de la connaissance de soi", (2000), *Philosophiques,* 27, 1, 161-186.
Proust, J. (to appear). Does metacognition necessarily involve metarepresentation *? Behavioral and Brain Sciences.*
Reid, T., *Essays in the Intellectual Powers of Man,* Essai III, ch. 4 et 6 ; reproduced in Perry, 1975.
Ricoeur, P. (1990), *Soi-même comme un autre,* Paris, Editions du Seuil. English transl. *Oneself as another*, Chicago, University of Chicago Press, 1992.
Shoemaker, S., & Swinburne R. (1984), *Personal Identity*, Oxford, Blackwell.
Shoemaker, S. (1970), Persons and their past, *American Philosophical Quaterly*, 7, 4, pp. 269-285. reproduced in *Identity, Cause and Mind,* Cambridge, Cambridge University Press, 1984, pp. 19-47.
Shoemaker, S. (1996), *The First-Person Perspective and Other Essays*, Cambridge, Cambridge University Press.
Stephens G.L. & Graham, G. (2000), *When self-consciousness breaks*, Cambridge, Mas. : MIT Press.
Velleman, J.D., (1996), Self to self, *The Philosophical Review*, vol 105, 1, 39-76.
Velleman, 2002
Williams, B., (1973), *Problems of the Self,* Cambridge, Cambridge University Press.