To appear in: Jesús Aguilar & Andrei A. Buckareff (eds.), *Philosophy of Action: 5 Questions*, Automatic Press/VIP

My answers to five questions on agency

Joëlle Proust (Institut Jean-Nicod)
EHESS & ENS, CNRS, Paris

1. Why were you initially drawn to theorizing about action and agency?

I had been introduced to the domain by John Searle's seminar on the philosophy of action, whose contents inspired his 1983 book entitled *Intentionality*. John is an impressive lecturer and on-line thinker; his seminar was, in two respects, deeply innovative. First, John claimed that action itself (rather than the beliefs and desires that might contribute to causing it) has a general structure common to all the Intentional states. Second, he introduced the famous notion of an intention-in-action, thus allowing to analyze as intentional all the spontaneous, embodied ways in which actions are performed. The combination of these two remarkable features of Searle's theory opened up brand new ways of thinking about embodied intentions, and would prove very useful for describing psychological and neurological evidence. My personal involvement in the philosophy of action, however, did not emerge until ten years later, when I set out to explore schizophrenic perturbations of agency awareness.

At the beginning of the nineties, I was conducting research on such general philosophical topics as intentionality, consciousness and concept acquisition, as a member of an interdisciplinary CNRS unit, CREA, where rich interactions between philosophers of mind and language, logicians, anthropologists, experimental psychologists, and psychiatrists were a regular feature. Given my interest in consciousness, I engaged in a collaborative research project with a French psychiatrist, Dr. Henri Grivois, a specialist in the early episodes of psychosis. According to him, patients generally have, at the onset of their illness, a sudden impression of enjoying a deeply felt, reciprocal bond with others, which Grivois named "concernment". Subjects seem to think that people (and objects) convey messages to or about themselves. After a few hours or days, patients come to believe that they have a specific role to play in the community, or even in the world at large – thus arriving at an experience of "centrality". The second group of features in Grivois' clinical description relates to a

perturbed agency. Patients have the impression that others are all doing the same thing, and are mainly imitating them; they however are the ones who imitate others - either in gestures (echopraxia) or in words (echolalia). They also frequently emulate others' goals. These two group of features, concernment and mimetic behaviors, Grivois speculated, are functionally related. Patients might feel concernment and centrality because of their mimetic disposition.

As a philosopher interested in consciousness, I found that Grivois' theory was quite intriguing. It helped one to approach consciousness from a fresh perspective, based on the direct functional requirements of agency. Dan Dennett, in his *Consciousness Explained*, defends a 'Multiple Drafts model' of consciousness that also tries to assimilate scientific evidence. On his view, information entering the nervous system in a massively parallel way, is locally 'edited' by the brain, with different quasi-narratives emerging at different time scales. Dan was then mainly interested in visual perception and memory; an efficient general rebuttal of the 'Cartesian theater' view of consciousness, his model does not say much about the specific roles of conscious states in action. In a footnote (p.112), however, Dennett briefly reports a 1963 experiment by T.I. Nielsen having to do with agency awareness. A subject is tricked into believing that he is looking at his own gloved hand when drawing a figure : the perceptual 'reafference' for his action – what he actually looks at - is sometimes his, sometimes a hand that performs a different movement. Placed in this condition, a subject who is visually monitoring his drawing hand feels it being *pulled* in the direction of the observed movement each time the observed movement conflicts with his actual performance. In Dennett's terms, this particular brain 'editing' or 'inference' solves a conflict between vision and proprioception of one's actions. How can an inference be experienced as a pull?

The parallel of this case with the feeling of 'concernment' of Grivois' patients in their first psychotic episode is striking. In both cases, subjects have a proprioceptive feeling 'telling them' that something has interfered with their will. In both cases, some kind of conflict is involved between what the subjects predict and what they observe about the consequences of their acting. Grivois'observations contained the seeds of a view, developed since then, that a non-matching comparator might trigger the psychotic sense of extraneity – of a foreign interference - for one's actions and thoughts. An important difference between the two cases, however, is the following: where Nielsen's subjects felt a strong pull to the direction of the (seen) movement, patients with psychosis would feel a diminished sense of agency for their own actions. Where Nielsen's subjects were exposed to an occurrent (accurate) non-matching signal thanks to a functional comparator, patients were exposed either to a permanent (inaccurate) non-matching signal, or to no signal at all, because of a faulty comparator.

This parallel, among other considerations, motivated us to start testing experimentally the 'motor' hypothesis in patients with schizophrenic delusions, with the collaboration of Marc Jeannerod, a neuroscientist of action. If we put a deluded patient in an experimental situation such as Nielsen's, will he be able to tell whether the currently observed hand is his own? We predicted that patients should be impaired, relative to non-deluded patients and normal controls, in recognizing their own actions. Our predictions turned out to be correct (Daprati et al., 1997). Similar conclusions were reached by Chris Frith and his colleagues, using different experimental paradigms and stimuli. Now, the important question is: what philosophical benefit can be drawn from this empirical research?

2. What do you consider to be your own most important contribution(s) to theorizing about action and agency, and why?

The most direct effect of interdisciplinary influences on my philosophical work was to press me into articulating a metaphysics of action that could reflect the functional aspects that psychologists and neuroscientists have been uncovering. Part of what clinical evidence taught us was that the senses of bodily and of mental agency are simultaneously perturbed in deluded subjects: their actions seem to them to be externally controlled, and their thoughts to be inserted. This suggests that an adequate theory of action should apply to both object- and self-directed kinds of action. Clearly, action is a dynamic phenomenon through which an agent uses representations to control her physical and social environments or her own thought processes. To reflect this dynamic property of actions, a proper definition needs to express the teleological structure of the causal-representational process through which a particular goal (and a trajectory to reach it) is selected among a range of actions in the agent's repertoire. How are we to accommodate these various functional requirements in defining an empirically sound concept of action?

A first revision is now well-accepted. Beliefs, desires and intentions, with a propositional content, as ordinarily conceived, do not deal easily with the specific ways chosen to perform an action. Granting that agents have preferential ways to produce outcomes, intentions in action must be able to have *nonconceptual*, that is, protopropositional kinds of content. These contents specify the spatial and temporal dynamics, rhythm, proprioceptive markers, and other phenomenological properties that have to be present for the

action to be successful (Proust, 2003). Given that these nonconceptual contents drive motor activity, they are usually called "motor representations" - a term used by neuroscientists to refer to neural activations in the motor area. Motor representations, however, only apply to bodily action. When applied to thought processes, other types of nonconceptual content, with their own dynamic features, are used to select and monitor mental actions; these 'epistemic feelings' are particularly interesting to study, as they form a primary basis for self-knowledge. (Proust, 2007, 2008). A second step was to examine the requirements for the representation of an intention to acquire *executive force*. Not every candidate intention to act is executed; intentions can also be simulated, evaluated and abandoned. How should one distinguish the executive from the simply motivational and epistemic activity evoked by an intention to act? A volitional definition of action was developed in Proust (2005) to spell out this difference.

This definition has a teleological structure, which in the case of action is roughly expressed by the law of action-effect: Actions are behaviors that tend to be executed in a context, and they tend to be executed in this context because they have had a beneficial effect for the agent in this context. Behaviors that have had detrimental consequences in this context tend not to be reproduced. This general teleogical structure needs to be complemented, however, in order to 1) express the dispositional range of accessible actions for an agent at t; 2) explain why unsuccessful or inadequately targeted actions may occur; and 3) explain how action patterns can, on a particular occasion, be adjusted and modified to respond to the constraints of a new context. The mathematics of adaptive control systems is helpful in satisfying these requirements with full generality. Regulation laws state which outcomes are associated with specific commands in specific contexts. Feedback laws state which portion of the regulation space is accessible at a given time to an organism with a given learning history. Regulation and feedback laws have been shown to necessarily involve *representations of the action space* (Proust, 2006a). It is indeed a theorem of control theory that the best regulator of a system is one which is a model of that system. In an optimal control system, in other words, the regulator's actions are, as Conant and Ashby write, "merely the system's actions as seen through a specific mapping".[1] In other words, any regulation involves a prior simulation. Action needs to be accounted for in a way that combines such dynamical and representational properties. The concept of volition to act, which I articulated in 2005, attempted to provide such an account. A volition is a dynamic executive event, that decomposes into the following three interconnected properties:

[1] Conant & Ashby, 1970

1) A regulation space is available to the agent at the relevant time, t, in which there is at least one trajectory that leads from the current position to a target;
2) There is a salient target at t;
3) A prevailing reward associated with the salient target selects and activates a suitable command at t.

Note that these three conditions correspond respectively to procedural-epistemic, contextual and motivational conditions. They have to apply jointly for a volition to occur (they are necessary and sufficient conditions for a representation of the goal to be endowed with executive force). On this view, a volition does not constitute a self-contained mental action that would prepare bodily action. A volition to bring about G is the representational-executive contribution that normally generates the effects typical of G (whether mental or bodily). Note also that volitions for bodily as well as for mental actions are made accessible by the system's prior history. As in any teleogical-etiological definition, such history explains why volitions exist, and how the parameters involved in the three conditions are properly calibrated, or, in certain biased cases, distorted. The particular history profile of an agent is thus able to explain how she can mistakenly represent that a certain kind of control will bring about a given outcome.

In a (non-dynamic) semantic analysis of volition, a volition is defined as an effort to bring about some change *in virtue of this effort*. One of the benefits of articulating the control structure above is that the *reflexivity* of volitions, intentions, and reasons to act can now be understood as an architectural feature derived from the control structure. As we have seen, control systems are essentially causal-representational loops. In each loop, a given command causally determines, in virtue of its content, a corresponding episode of monitoring, which in turn triggers, in virtue of its content, a new command: reinforcing the former command, or prompting revisions to it. The property of mutual causal influence between these two complementary levels is called "causal contiguity". (Mellor, 1991) Now, an important semantic fact is that *any* causally contiguous representational structure necessarily generates both *representational promiscuity* and a basic, nonconceptual form of *reflexivity*. (Proust, 2007).

There is representational promiscuity because command and monitoring need to compare observed feedback with expected feedback; they need to have a common representation of the goal, and of the route that leads to it. There is reflexivity because monitoring always refers to the specific command that prompts it. Both representational promiscuity and reflexivity have been noticed by classical analyses of action. If you intend to

act, your intention also constitutes the satisfaction condition of the completed action: this representational promiscuity indeed depends on the causal/semantic relationship between a command and a monitoring episodes. They only differ in that the first states what is to be obtained, while the second reports on what does/does not obtain (a difference in direction of fit,[2] with the same basic content). In every cognitive control structure, representational promiscuity must be present. Monitoring *evaluates* the execution of the command, while the command *directs* or organizes upcoming monitoring. In contrast with beliefs, desires, and intentions, volitions thus involve two directions of fit, based on the same representational schema (as suggested by the idea of comparing what you have with what you want).

Let us now examine how volitions to act have reflexivity as a consequence of representational promiscuity. Reflexivity is necessarily generated in a causally contiguous representational loop for two reasons. Resulting as it does from a causally contiguous command episode, every monitoring episode *carries information* about the previous command that caused it. Furthermore, some representational element in monitoring must have the function of carrying this information. It is crucial indeed that the command level identifies the ouput of a monitoring episode as a response to its own query, which requires in turn that the content of monitoring is structured by the command that causes it. The representational element having the function of a *de se* marker, however, does not need to involve a self-concept. (Proust, 2003). This result will not be surprising to those philosophers who, like John Perry, claim that "basic self-knowledge is intrinsically selfless" (1986, 2000). As Perry has shown, it is possible to have information concerning something without representing it explicitly in the thought content. In that case, reflexivity is mentally by an "unarticulated" constituent. Recanati (2007) argues that cases of identification-free reflexivity (those that are immune to error of identification) depend on the mode of thinking (like perceiving or intending) rather than on the content of thought. On my own view, the "reflexive" modes are those that involve causal contiguity between two representational events: a (context-sensitive) command episode and a (circumstance-sensitive) evaluation episode (i.e. monitoring).

Obviously, these philosophical observations are relevant to understanding the schizophrenic delusions described above. Adopting a control view of action in fact allows us

---

[2] A belief is true when it fits the world : it has a mind-to-world direction of fit. Reciprocally, a desire, an intention or a volition are satisfied if the world adjusts to them : they have a world-to-mind direction of fit. See Austin, 1962.

to adopt a new perspective both on agency and on agency awareness. Recently, philosophers have tried to explain the asymmetry between a preserved sense of ownership (of bodily experience) and a perturbed sense of agency in patients with schizophrenic delusions of control. On the present view, there are as many forms of reflexivity and possible disturbances as there are control loops. Functional considerations suggest that there are at least four such levels, which are respectively *motor* (intentions in action/volitions to move), *self-world* (prior intention/volition to obtain that P- a given change in the world), *self-self* (prior intention/volition to obtain a mental property by affecting oneself), and *self-other* regulations (prior intention/volition to obtain that P in virtue of our common effort). Abnormal phenomenology results from reflexive blindness generated in one or the other of these loops. Reflexive blindness, however, belongs to monitoring, rather than control. It does not affect to the same degree the capacity to act. I tried to explain this interesting dissociation in schizophrenia in Proust (2006b).

3. What other sub-disciplines in philosophy and non-philosophical disciplines stand to benefit the most from philosophical work on the nature of action and agency, and how might such engagement be accomplished?

To understand how an empirically adequate conception of action can gain higher relevance to other fields, one obviously needs to have some idea of the features of the philosophical theory that will fulfill this promise. As I have said above, my own view is that the most empirically adequate way of capturing the conceptual structure of action is to see it as the most accomplished form of adaptive control to date. Adaptive control is itself an adaptation whose vehicle is constituted by embodied neural systems. The kind of adaptive control a being possesses determines, by and large, the framework within which his/her perception, emotion and memory will play their respective roles. If one adopts this view, the four most directly affected fields are, in science, mathematics, evolutionary biology, and the cognitive sciences (including cognitive psychopathology), and, in philosophy, epistemology.

Why *mathematics*? Mathematics is close to philosophy in that it aims to capture necessary relations between propositions. Dynamic system theories, such as the viability theory, starting from the hypothesis that dynamic systems are associated with changing environments, allow us to understand and describe the constraints to which agency needs to respond. In particular, viability theory, (Aubin, 1991) allows the modeling of rational decisions in uncertain contexts in a way that radically minimizes the number of

parameters that need to be represented for a preference to emerge. If the metaphysical assumption that agents are a subclass of adaptive control systems is correct, mathematicians, philosophers and cognitive psychologists will have to work together in order to come up with specific classes of control models of action that meet the crucial constraint of parametric reduction.

Evolutionary Biology studies the interaction between evolutionary history, adaptations, and selection pressures, and, in particular, the evolution of mental capacities across phylogenetic trees. It can be fertilized by philosophical work on agency, as is shown by the impact on the field of Sterelny's (2003) book. Much more needs to be done, however, to grasp the evolution of forms of action in non-human animals and the role which the linguistic expression of goals and values can play in controlled agency.

The Cognitive Sciences, and in particular Cognitive Psychopathology and Neuropsychology, have already drawn inspiration from the philosophy of action, by using, sometimes uncritically, concepts able to describe behavioral evidence (for example, the notion of an intention-in-action has been widely used by neuroscientists.) We have seen, furthermore, how philosophy and psychiatry have influenced each other in determining the dimensions of self-awareness involved in agency, and worked jointly with cognitive scientists on accounts of psychiatric disorders.

Epistemology, however, strikes me as the philosophical subdomain that has the most to gain from current research on mental agency. In order to fully understand the concept of entitlement or prima facie justification, in which a subject feels entitled to form a belief without being able to explicitly articulate the reasons that he has to form that belief or the norm that he is applying, one needs to explain philosophically how this implicit knowledge has been formed and what content it has. Current work on mental action, and, in particular, on the control structure that it involves (in metacognitive judgments such as: *I need to stop/pursue this line of reasoning*, or in an attempt to retrieve an item from memory) is able to shed light on this issue. (Proust, 2008)

4. What do you regard as the most neglected issues in contemporary work on action and agency that deserve more attention?

Relying as they mostly do on folk-psychological intuitions, analytic philosophers tend to conflate two dimensions of agency: the motivational and the executive. Distinguishing volition from intention, is meant to provide the required distinction : you can desire A, know how to get A by doing P, intend to do P to get A, while failing to attempt to do P.

Conversely, you may execute P without having a reason, or a motivation to do so. As psychologists such as Theodule Ribot have shown us, there can be neuropsychological disturbances of volition that do not affect motivation, and conversely (see the imitation syndrome, described in Lhermitte, 1986).

A movement that is voluntarily performed feels distinctively "effortful", and the effects of a willful movement have a distinctive temporal signature as compared with effects passively observed (Haggard, 2003). Volition itself, however, has no specific phenomenology; Malebranche correctly stressed that agents start moving their body (or mentally focusing, in mental action) to attain their goals, without knowing how they do it. My own account as summarized above only offers a teleological explanation for the switch from 'epistemic-motivational' to 'active'; this type of definition identifies a selected function without offering a detailed metaphysical solution. The puzzle that a teleological explanation solves is that agents cannot be said to voluntarily switch into the active condition. In some sense, willing is something that happens to you, which does not mean that you cannot select from among desires those that you want to influence your future willing episodes. It strikes me, however, that the issue deserves to be dealt with in more detail, by studying the properties of dynamic systems, or otherwise.

5. What are the most important open problems in philosophical theorizing about action and agency, and what are the prospects for progress?

Philosophers enjoy doing mostly what they know how to do; as a result, agency has up to now mainly been studied in philosophy through attitudes and their propositional contents. The domain of emotion, however, has proved difficult to approach in these terms; similarly, the domain of action resists a simple belief-desire analysis. In spite of these well-known problems, most analytic philosophers stick to their guns. Why should they have to understand brain activity to theorize about agency, or about consciousness? The less science, the more philosophy (the saying goes)! Is not a proper distinction between a personal level (what I feel and can report about my states of awareness) and subpersonal levels (what happens in my brain that causally explains how I feel, report, and do things) able to justify us in *not* bothering to collect scientific evidence?

The problem with this general strategy is that the distinction that it presupposes is highly variable, self-interpretive and theory-laden. As Dennett took pains to show, the limit between subpersonal and personal facts may change with time scales, attentional demands and health. There is no firm ground in the personal that is not constrained by a myriad of

subpersonal representations, which philosophers cannot afford to entirely ignore, in particular if they recognize the existence of nonconceptual forms of mental content. The mental cannot reside in a supposedly 'final' narrative layer, as articulated in language. It seems difficult, in short, to deny that neuroscience, experimental psychology, neuropsychology and cognitive psychopathology deliver facts and evidence that any conceptual view of consciousness or agency needs to take into account if it is to be empirically adequate.

This does not mean, however, that philosophers need to summarize endless facts about the way the brain sends action commands and monitors them. What is rather needed, is serious conceptual work to grasp the dynamic properties of agentive systems, and to come up with views on how these possibly non-conceptual representations might be conditions of action selection and recognition. On my view, the prospects for philosophical progress on this issue presently call for the collaborative insights of mathematicians and computer scientists, on the one hand, and philosophers and psychologists, on the other hand.

Acknowledgment

I express all my thanks to my colleague Dick Carter for his comments and his linguistic help.

References

Aubin, J.P. 1991. *Viability theory*. Boston, Basel : Birkhauser.
Austin, J.L. 1962. *How to do things with Words: The William James Lectures delivered at Harvard University in 1955*. Ed. J. O. Urmson. Oxford: Clarendon.
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: 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, 71- 86.
Dennett, D. 1991. *Consciousness Explained*. Boston, Little, Brown & Company.
Frith C.D., (1992), *The Cognitive Neuropsychology of Schizophrenia*, Hillsdale, Lawrence Erlbaum Associates.
Grivois, H. 1998. Coordination et subjectivité dans la psychose naissante. In H. Grivois et J. Proust (eds.), *Subjectivité et conscience d'agir, approches clinique et cognitive de la psychose*, Paris, Presses Universitaires de France, 1998, 35-73.
Haggard, P., 2003. Conscious awareness of intention and of action, in J. Roessler & N.Eilan (eds.), *Agency and self-awareness: issues in philosophy and psychology*, Oxford, Oxford University Press, 111-127
Lhermitte, F., Pillon, B., Serdaru, M.1986. Human Autonomy and the Frontal Lobes. Part I: Imitation and Utilization Behaviour, *Annals of Neurology, 19, 4,* 326-334.
Mellor, H. 1991, I and now, in: *Matters of Metaphysics* , Cambridge, Cambridge University Press, 17-29.
Nielsen, T.I. 1963.Volition: A new experimental approach, *Scandinavian Journal of Psychology,* 4, 225-230.
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*, 129, 105-128
Proust, J. 2003_a. Action, in B. Smith (ed*.), John Searle,* Cambridge, Mass.: Cambridge University Press, 102-127.
Proust, J. 2003_b. Thinking of oneself as the same, *Consciousness and Cognition*, 12, 4, 495-509.

Proust, J. 2005. *La nature de la volonté,* Paris, Folio-Gallimard.
Proust, J. 2006$_a$. Agency in schizophrenics from a control theory viewpoint, in W. Prinz & N. Sebanz (eds.), *Disorders of volition*, Cambridge, MIT Press, 87-118.
Proust, J. 2006$_b$. Rationality and metacognition in non-human animals, in S. Hurley & M. Nudds (eds.), *Rational Animals*?, Oxford, Oxford University Press, 247-274.
Proust, J. 2007. Metacognition and metarepresentation : is a self-directed theory of mind a precondition for metacognition ? *Synthese,* 159, 271-295.
Proust, J. 2008 (in print). Is there a sense of agency for thought? In L. O'Brien ed. , *Mental action*, Oxford, Oxford University Press.
Searle, J.R..1983. *Intentionality, an Essay in the Philosophy of Mind,* Cambridge, Cambridge University Press
Sterelny, K. 2003. *Thought in a Hostile World, The evolution of Human Cognition,* Oxford: Blackwell.