

COGNITIVE SEMIOTICS

ISSUE 0 (SPRING 2007)

AGENCY

EDITED BY: LARS ANDREASSEN, LINE BRANDT & JES VANG

-
- 3 What is Cognitive Semiotics? A general introduction to the
journal
- 5 Editorial Preface
- 8 Søren Overgaard & Thor Grünbaum
What Do Weather Watchers See? Perceptual intentionality and agency
- 32 Shaun Gallagher
Sense of Agency and Higher-Order Cognition: Levels of explanation for
schizophrenia
- 49 Rick Grush
Agency, Emulation and Other Minds
- 68 Merlin Donald & Lars Andreassen
Consciousness and Governance: From embodiment to enculturation - an
interview
- 84 Kristian Tylén
When Agents Become Expressive: A theory of semiotic agency
- 102 Mikkel Holm Sørensen & Tom Ziemke
Agents without Agency?

COGNITIVE SEMIOTICS

EXECUTIVE EDITORS

Line Brandt, Per Aage Brandt, Frank Kjørup, Todd Oakley, Jacob Orquin, Jakob Simonsen, and Jes Vang

FINANCIAL DIRECTOR

Per Aage Brandt, Case Western Reserve University, Department of Cognitive Science, 10900 Euclid Avenue, 44106-7179 Cleveland, Ohio, USA. Tel.: +1 216 368 2725. Email: pab@cognitivesemiotics.com

EDITORIAL MANAGER (address for submissions and general editorial contact)

Jes Vang. Email: jv@cognitivesemiotics.com

WEBMASTER

Jacob Orquin. Email: jo@cognitivesemiotics.com

INTERNATIONAL ADVISORY BOARD

Peer F. Bundgaard, Jean Petitot, Frederik Stjernfelt, Wolfgang Wildgen, and Svend Østergaard

MANUSCRIPT SUBMISSIONS

For style guide and other directions for authors go to the journal's website: www.cognitivesemiotics.com

PUBLISHING DETAILS

Cognitive Semiotics 0 (Spring 2007) is published as a special promotional issue. All copyrights owned by the executive editors.

SUBSCRIPTION AND ORDERING INFORMATION

Cognitive Semiotics 1 (Autumn 2007) and following issues of the journal will be internationally published by:

Peter Lang AG
European Academic Publishers
Moosstrasse 1 · P.O. Box 350
CH-2542 Pieterlen · Switzerland
Tel.+41 (0)32 376 17 17
Fax+41 (0)32 376 17 27
e-mail: info@peterlang.com
Internet: www.peterlang.com

- See Order Form at the back of this issue

PRINTED BY

Werk Offset A/S, Bødstrupvej 2-4, DK-8270 Højbjerg, Denmark. Tel.: +45 8627 5499. Fax: +45 8736 0723.
Email: werk@werk.dk

FINAL PROOF READING

Jiaxi Ding and Max Jensen

COVER DESIGN AND GRAPHICAL LAYOUT

Jes Vang

SPECIAL THANKS TO

Martina Fierz at Peter Lang AG for her invaluable and kind assistance

What is Cognitive Semiotics?

A general introduction to the journal

We are pleased to present this very first issue of *Cognitive Semiotics*. The journal will publish two print issues a year, one in the spring and one in the fall. In addition, novel content will be made available periodically at the journal's website which also serves as a vital resource for the journal.

As a whole *Cognitive Semiotics* offers its readers the opportunity to engage with ideas from the European and American traditions of cognitive science and semiotics, and to follow developments in the study of *meaning* – both in a cognitive and in a semiotic sense – as they unfold internationally. The intention of the journal is to create and facilitate dialogue, debate, and collaboration among those interested both in human cognition and in human semiotic experiences and behavior.

This intention is inherited from its Danish antecedent, the journal *Almen Semiotik*, published by the Aarhus University Press (1990 – 2007). The initiative to create a transatlantically based journal comes from the Center for Cognition and Culture, at the department of Cognitive Science at Case Western Reserve University, and from a group of researchers trained at the Center for Semiotics, in Denmark, and based in Aarhus and Copenhagen. These joint editors identify the present issue as issue number “0” to signify its transitional status. We are happy that *Cognitive Semiotics* will be published by Peter Lang AG, where the book series *European Semiotics*, created in 1999, is also housed.

Let us briefly explain the general content of this journal, the field of thinking and research we name *cognitive semiotics*.

Human minds ‘cognize’ and ‘signify’ as complementary aspects of their capacity to think and feel. If we accept the metaphor of ‘higher’ and ‘lower’ levels of cognition, and the idea of seeing the ‘higher levels of cognition’ as those responsible for abstraction, language, discourse, institutions, law, science, music, visual arts, and cultural practices in general, grounded in the use of conventionally established and intentionally used signs (often called symbols), then semiotics is the discipline

committed to the study of these ‘higher levels’. Relying predominantly on expression-based communication, the contents of these higher-level cognitive feats can be shared by expressive exchanges of signified meanings (GE: *Bedeutungen*; FR: *sens*). These meanings, in turn, can be made the subject of inquiry, their semiotic structure and significance indicators of how minds cognize together, and of the cognitive mechanisms which make their production and comprehension possible in the first place.

The mental activities of thinking and communicating are importantly interrelated in our species. Human societies and cultures, and civilization at large, are the results of cooperating and conflicting minds, connected through cognitive-semiotic functions and processes. To gain scientific knowledge about these often still unexplored phenomena, found increasingly important by the scientific community, the journal is devoted to high quality research, integrating methods and theories developed in the disciplines of cognitive science with methods and theories developed in semiotics and the humanities, with the ultimate aim of providing new insights into the realm of human meaning production and the modalities of its embodiment and disembodiment.

Cognitive Semiotics will publish peer-reviewed manuscripts according to a double-blind protocol. We invite authors to submit manuscripts on the above-mentioned and related topics to the managing editor at the address given on the masthead. Also, we encourage everyone to visit our website www.cognitivesemiotics.com for relevant updates and news, and to sign up for our electronic newsletter to be informed of the upcoming editions of the journal.

Editorial Preface

A recurring format for *Cognitive Semiotics* will be the appearance of select articles presenting different takes on a shared topic. This first issue of our journal has *Agency* as its special theme. The term Agency covers a broad spectrum of different theoretical approaches ranging from the phenomenologically inspired idea of an active intentional subject interacting with other agentive subjects and responding to the affordances of objects appearing in the physical environment (e.g. their salient features and possibilities for sensorimotoric interaction) to a (neo)behaviourist concept of Agency concerned with computer-simulated networks of “Agents” reacting within a “System” of inhibitors, e.g. physical or social restrictions.

The variety of notions and uses of the term “agency” pose an apparent stumbling block for a proper synthesis. However, if we concentrate on the semiotic question at hand – what ‘agency’ *means* – we might be able to get closer to some distinguishing, perhaps even universal, qualities of the phenomenon and its intrinsic human value for us as cognitively distributed and embodied subjects.

Several notable features seem to reappear across the theoretical platforms. One is the tight link between the “sense of agency” and kinaesthesia, and concerns our ability to initiate self-movement – actual or hypothetical – which contemporary research has shown serves as a precondition for having proper perceptual experiences, complete with intersubjective, intentional structure. This line of argument is pursued by **Søren Overgaard** and **Thor Grünbaum** who maintain, with help from Husserl, that genuine visual perception of spatial objects originates in the subject being able to engage in ‘free kinaesthetic movements’, and that this ability is necessarily tied to, if not responsible for, the sense of agency. The sense of agency is then understood as an implicit awareness of *what* one’s body is doing in functional relation to *who* is doing it, i.e. one’s self as the instigator of intentional movement.

While the definition above might seem intuitive and straightforward, **Shaun Gallagher** delves deeper into its neurobiological premise and shows that for schizophrenic patients the *mis*attribution of agency to other people or events might in fact be due to the disruption of overlapping neurological components responsible for

representing Self and Other. In other words, this “Who” system is an integral part of the sense of self-agency and perhaps even the seat of basic intersubjective understanding, as the same brain areas are active when the subject engages in a particular intentional act and when he watches the act performed by someone else.

Also in search of neurobiological grounds for agency, **Rick Grush** ventures into the realm of mental representations of the self and the body. Arguing from the perspective of his *Emulation Theory of Representation* he makes the case that sophisticated neural ‘emulators’ of subject-object interaction, evolved from processes aimed at obtaining better motor control, are central to our understanding of ourselves as agents, and to imagining and correctly anticipating the outcome of our self-propelled actions (as well as the actions of others).

Taking a step towards a bigger picture the discussion between **Merlin Donald** and **Lars Andreassen** adequately sums up the arguments advanced by the previous authors while also positioning the phenomenon of agency within a broader human and cultural context. Thus, the authors present an analysis of the intricate social worlds in which we navigate and the complex cultural artifacts that we use for externalizing and augmenting our cognitive abilities, e.g. our memory systems.

In a similar vein, social gestures and cultural relics are examined when **Kristian Tylén** turns his semiotic attention to how meaning is derived from seemingly inconspicuous phenomena. We perceive the effects of intentional signification when interpreting the causal agency behind an erected headstone or an alluring wink of an eye. It is an inherent semiotic ability originating in shared attention and in this way apparently a form of interpretive cognition exclusive to human beings.

The latter might explain why it has proven immensely difficult for artificial intelligence research and robotics to get a handle on true agency. You might be able to synthesize organisms that can be perceived as agents, but they will lack proper (emergent) agency as long as the notion of agency within the artificial intelligence community remains as fuzzy as the case is today. At least this is the claim made by **Mikkel Holm Sørensen** and **Tom Ziemke** in their review and instructive critique of contemporary research on complex self-organizing adaptive systems.

Altogether, these six articles point to a notion of agency as one of the most fundamental aspects to human experience. It is what defines us as individuals with distinctive *selves* and what provides us with the foundation for intersubjective

understanding and interaction. We hope you will enjoy the comprehensive examination of the topic presented within this special, inaugural issue of *Cognitive Semiotics*.

Happy reading!

The editors

Next issue (Autumn 2007) will bear the title *CONSCIOUSNESS & SEMIOSIS*. New articles by Liliana Albertazzi, Bernard Baars, Per Aage Brandt, Terrence Deacon, Anthony Jack, Jean-Luc Petit, Ernst Pöppel, Frederik Stjernfelt, Patrizia Violi, and Svend Østergaard.

Visit www.cognitivesemiotics.com for news and updates.

Søren Overgaard and Thor Grünbaum

What Do Weather Watchers See? Perceptual intentionality and agency

This paper defends a twofold thesis. First, contra Galen Strawson's recent claims, we argue that a subject's ability to perceive spatial objects essentially presupposes that the subject has an experience of her own voluntary movements. Second, we argue that while there is good reason to say that this subjective movement is a kind of agency, it would, pace certain other authors, be wrong to use the term "action" for it.

Introduction

It is widely agreed among phenomenologists that there is some kind of essential link between perceptual intentionality – the ability to have perceptual experiences that are experiences *of* something – and embodiment. Thoughts along these lines are found in Husserl, Sartre, Merleau-Ponty, and even, in some form, Heidegger. Moreover, most phenomenologists would argue that there is an essential link between perceptual intentionality and some kind of bodily agency. More precisely, they would claim that bodily movement, the ability to move oneself, is a condition for the possibility of perception, among other things.¹ Some of these phenomenologists also make the

¹ For a strong, recent formulation of such an argument, see Sheets-Johnstone (1999). An interesting fact that we can only note in passing here is that many phenomenologists seem considerably less interested in the question of the extent to which perceptual intentionality might in turn be a precondition for agency. One exception is Smith (1992).

further claim that to move oneself is to perform an action, and they consequently argue that perceptual intentionality depends on an ability to perform actions.²

In the present paper, we wish to discuss the question of the possible link between the ability to move oneself and visual perceptual intentionality. Our aim is twofold. First, we want to demonstrate that perception does in fact necessarily depend on the subject's experience of her own voluntary movements; this is the aim of the first two sections of the paper. Second, we want to investigate how this ability to move oneself is best described; we will approach this issue in the third part of the paper. We will argue that while there is good reason to say that the subjective movement that visual perception presupposes is a kind of agency, it would be wrong to use the term "action" for it.

The Weather Watchers

Not all philosophers agree that there is an essential connection between perceptual intentionality and agency. A good example of a conflicting perspective is found in Galen Strawson's book *Mental Reality*. In the centre of Strawson's book stands his attempt to refute the view which he labels "neobehaviourism". "Neobehaviourism", Strawson says, is "the view that mental life is linked to behaviour in such a way that reference to behaviour enters essentially into any adequate account of the nature of almost all, if not all, mental states and occurrences" (Strawson 1994: xi). According to Strawson, one way to undermine the neobehaviourist view is to imagine a race of creatures, "the Weather Watchers", who have a rich mental life, but lack any ability or disposition to engage in any type of observable behaviour. Weather Watchers are supposed to have sensations, emotions, desires, and thoughts, just like we do, but at the same time they are supposed to be "constitutionally incapable of any sort of behaviour, as this is ordinarily understood" (Strawson 1994: 251). In fact, Weather Watchers are "rooted", like trees, and completely unable to move, but their mental life is nevertheless strikingly similar to ours:

A Weather Watcher lives the rooted life, but there are many respects in which its mental life is like ours. It sees the sky and hopes the clouds are bringing rain. It watches a seed lodge in a gap between two rocks by the edge of the river. It forms

² Apparently, this claim is made by Sartre (1943: 359), Mulligan (1995: 204), and Zahavi (1999: 93). However, in some cases the claim may be only apparent and due mostly to terminological carelessness.

the belief that a tree may grow there before long, and hopes that it will. (Strawson 1994: 255)

Clearly, if what Strawson calls “neobehaviourism” is true, there cannot be such creatures. Even though we are not committed to a neobehaviourist view, our position is also at odds with Strawson’s description. If we are right that there is an essential connection between the capacity for bodily agency and the ability to have visual perceptions, then the Weather Watchers are impossible. Thus, this is the challenge that Strawson’s Weather Watchers present us with: we must be able to show that such immobile creatures would be unable to have visual perceptions in anything like the ordinary sense; that is, perceptions of spatial objects, such as rocks, trees, seeds, and riverbanks. We intend to meet this challenge, and will thus argue that Weather Watchers, considered as immobile from start to finish, cannot watch seeds lodge in gaps between rocks.

First of all, however, we must make clear the terms of the discussion. We have to critically examine Strawson’s adoption of Hume’s division of human knowledge. According to Hume, the objects of human knowledge “may naturally be divided into two kinds, to wit, *Relations of Ideas*, and *Matters of Fact*” (Hume 1993, Enq. IV, Part I: 15). Under the title “Relations of Ideas” we find the propositions of mathematics and logic, which are absolutely indubitable, but tell us nothing about the world, whereas knowledge about matters of fact, as based on sensory experience, informs us about the world, but falls short of absolute certainty. That $2+2$ should not equal 4 is impossible, whereas “[t]he contrary of every matter of fact is still possible” (ibid.). Throughout *Mental Reality*, Strawson employs this Humean division in roughly the following way. If Strawson wants to argue for some thesis – for example that there is no essential link between mind and behaviour – his strategy is to describe a hypothetical case of creatures that is in line with his thesis – e.g., creatures that have fully developed minds, but completely lack the ability to behave in any way. He then challenges his opponents to demonstrate that the conception of such creatures is *logically incoherent*. If the opponent cannot meet this challenge, Strawson feels entitled to conclude that there can be no *essential* link between mind and behaviour, although there might, as a matter of fact, be such a link as far as all creatures of planet earth, or even the universe, are concerned. The crucial thing is that a deviant case is not logically impossible, and this is enough to show that things *might have been otherwise*; i.e., whatever link there might be

between behaviour and mind it can only be contingent, not necessary or essential (cf. Strawson (1994: 253, 265).

But is there really nothing in between merely factual, contingent truths, and the analytical truths of logic and mathematics? Or, to put it differently, does it follow from the fact that we cannot prove that some hypothetical scenario is *logically* incoherent that we must grant the intelligibility and even possibility of such a scenario? We want to invoke stronger and more demanding notions of intelligibility and possibility than Strawson is prepared to grant.³ We want to suggest a distinction between merely *talking* about something and being able to actually *imagine* this something being the case. To speak the language of Husserl, there is a crucial difference between the “signitive” act of referring to something in merely talking about it, and the “intuitive” act of imagining something. The difference might perhaps be illustrated by the marked difference between the case where one is casually referring to some terrible event, say the murder of John F. Kennedy, and the case where one actually imagines the event in all its hideous details. The act of vivid imagination here gives the former, “empty” intention, intuitive “fulfilment”. The paradigm case of such a fulfilling act is of course the act of (visual, auditory, tactile, etc.) perception, in which the intended object is present “in flesh and blood”, as Husserl puts it. But it is nevertheless important to stress that vivid imagination can also function as a fulfilling act.

Let us take a somewhat different example. It is perfectly possible to talk about a “square circle”; we know what the words mean, and we can understand what the expression as a whole is supposed to mean – viz. a circle that has the property of being square. However, we cannot imagine what a square circle might look like, let alone perceive such a circle. In other words, we can have an “empty” intention of a “square circle”, but we cannot provide the concept with any intuitive fulfilment. Of course, the notion of a “square circle” is also logically incoherent – it is a textbook example of a contradiction. In order to bring out the difference between lack of intuitive fulfilment and logical incoherence, we must consider a third example. Most of us can imagine the

³ One might also have reservations about Strawson’s extensive use of thought experiments. Philosophers are sometimes a bit too confident that we are able to draw interesting conclusions from any hypothetical scenario that we can formulate intelligible sentences about. But is that necessarily so? Might such thought experiments not be so unusual and strange that we really do not know what to conclude from them? As Daniel Dennett once wrote, “[w]hen philosophical fantasies become too outlandish – involving time machines, say, or duplicate universes or infinitely powerful deceiving demons – we may wisely decline to conclude *anything* from them. Our conviction that we understand the issues involved may be unreliable, an illusion produced by the vividness of the fantasy” (Dennett 1982: 230). For a useful, brief discussion of these issues, cf. Parnas and Zahavi (2000: 7-11).

possibility of moving physical things from a distance with one’s mind alone. Although as far as we know, it is not actually possible to do this, we can make intuitive sense of the notion of being able, by an extreme effort of concentration, to move a glass on a table in this way. For instance, we can imagine that our mental concentration becomes a force that somehow pushes the glass, like our breath can move the flame of a candle. However, consider the following modification of our example: Imagine that you are able, again just by concentrating, to move the glass on the table; but this time you are not, as it were, moving the glass by the aid of something else (a mental force that functions like a hand). Rather, you are moving it (from a distance, with your mind alone) in the same direct way that you move your own limbs. It is not possible to form any clear intuitive notion of such a scenario. And yet, we understand the sentences describing it, and there seems to be no logical incoherence in the description. It is a situation, then, that we can talk about, but not really make sense of – one we cannot provide with any intuitive fulfilment.⁴

In *Ideen I*, Husserl refers in a similar context to scenarios that are logically possible yet “factually absurd”. “Factual absurdity” (*sachlicher Widersinn*), is Husserl’s term for something in between mere facts and strict logic. Contrary to what the phrase “factual absurdity” might tempt one to think, what is at stake *is* a kind of absurdity; it does not just concern “matters of fact”. Factually absurd scenarios, according to Husserl, are precisely scenarios that involve no logical contradiction, but which we cannot make sense of if we try to probe a bit deeper into their preconditions and implications (Hua III/1: 102).⁵ It is our claim that something like this is true of Strawson’s Weather Watchers. It is characteristic of Strawson’s use of hypothetical scenarios, including the one with the Weather Watchers, that he says relatively little about the various creatures he imagines. Instead he challenges his reader to prove them logically incoherent. But as we have tried to show in the present section this defence will not do. We must be able to “intuit” these hypothetical scenarios to understand and vividly imagine them – and we cannot do this in the case of the Weather Watchers. In the following sections, we try to substantiate this claim by advancing an argument for the thesis that visual perception essentially presupposes a capacity for bodily agency.

⁴ For further discussion of this scenario, see Danto (1973: 138-141).

⁵ We refer to the volumes of Husserl’s collected works (*Husserliana*) in the standard way: the abbreviation “Hua” followed by volume number (Roman numerals) and page number.

Horizon intentionality and kinaesthetic sensations

The question to be addressed in this section might be phrased in the following way: How can we make visual perception intelligible as perception of three-dimensional objects in three-dimensional space? More specifically, the question is whether one can make sense – “full” intuitive sense – of visual perception for creatures essentially incapable of any bodily action or movement.⁶

This question is one that has received attention from many phenomenologists. Most famous, perhaps, is Merleau-Ponty’s account in *Phenomenology of Perception*, but one might also mention Sartre’s chapter on the body in *Being and Nothingness*. However, the most sustained effort to come to grips with these issues is perhaps found in the manuscripts of Husserl. From 1907 and onwards until the end of his life Husserl worked intensively on the question of the connection between perceptual intentionality and bodily movement. For the sake of simplicity, we will focus on Husserl’s account in the following discussion.

In his phenomenological investigations of perceptual intentionality, Husserl most often uses familiar, everyday examples such as perceptions of houses and trees. When perceiving a tree, I typically see it as a tree in a garden surrounded by a lawn, or in a forest surrounded by other trees. In other words, a tree presents itself to me in certain “object-surroundings”; it does not normally manifest itself in complete isolation. Using Husserl’s idiom, we can say that a tree is perceived as surrounded by a perceptual “horizon” of other objects. This horizon even includes objects that are not momentarily perceptually presented. For example, when I look at a tree in a familiar park, the tree presents itself as part of a reality that stretches far beyond what is perceptually given to me; the rest of the park, and parts of the surrounding city are dimly co-intended as well. In fact, the perceptual horizon of this present act of perception is not even exhausted by these open-ended object-surroundings. According to Husserl, a perceptual horizon belongs to the perceived tree itself, independently of its surroundings. The tree presents itself as one that I could view from different angles,

⁶ We focus on visual perception in order to keep our discussion within reasonable bounds. We do not want to claim that the visual sense enjoys some privileged status in comparison with e.g. the tactile sense. Nor do we mean to suggest that one can completely separate the contributions of the various senses as if these were perfectly independent of each other. Rather, we recognize that the various senses are intimately connected and interwoven with each other. The necessary limitations of a journal article, however, demand that we simplify things slightly in the following discussion.

thereby bringing different sides (such as the presently “invisible” backside of the trunk) into view, and also as one that I can inspect more closely with regard to the actually perceived side itself (Hua XI: 7). These multiple different aspects of the tree are co-intended, according to Husserl, in my present perception of the tree, even though this perception, strictly speaking, only presents one such aspect:

The object is not actually given, it is not given wholly and entirely as that which it itself is. It is only given “from the front,” only “perspectivally foreshortened and projected” etc. While many of its properties are illustrated in the core content of the perception, at least in the manner which the last expressions indicate, many others are not present in the perception in such illustrated form: to be sure, the elements of the invisible rear side, the interior etc., are co-intended [*mitgemeint*] in more or less definite fashion [...], but they are not themselves part of the intuitive, i.e., of the [strictly] perceptual or imaginative content, of the perception. (Hua XIX/2: 589)

A spatial object essentially presents only a profile out of a multitude of possible profiles. It is important to emphasize that these additional profiles belong to the perceived object precisely as intended in the perception in question (Hua XVI: 50). To put it differently, it is *according to the concrete perception* itself that the object has “more to offer”. The intentional surplus must *itself* be considered a *perceptual* intentional surplus, rather than some non-perceptual intentionality, such as, for example, an imaginative intentional complex, always accompanying a perception (cf. Hua XVI: 56). What I perceive when I perceive a tree is not an aspect or a profile of a tree, nor (in most cases) merely one side of it, but precisely the tree itself, the “whole” tree, according to Husserl – and this means I perceptually co-intend much that is not itself, strictly speaking, manifest. The concrete perceptual manifestation encompasses both “proper” and “improper” manifestation (Hua XVI: 50), as Husserl puts it. Conversely, we may say that a *horizon* of co-intended profiles or aspects belongs to the perceived object just as perceived (Hua VI: 161). Thus, perception – the paradigm of an intentional experience that presents its object in flesh and blood – turns out to not present, in the strict or “proper” sense, its object exhaustively, after all. In Husserlian terminology, perceptual intentionality is “a complex of full and empty intentions” (Hua XVI: 57), since not all co-intended profiles are “saturated” with proper manifestation.

But the question, of course, is whether visual perception *must* have this “horizontal structure”. Husserl thinks it must. In *Ideen I*, he thus claims that even a god would have to perceive spatial objects this way if he were to perceive them *as spatial objects* at all (Hua III/1: 89). It is not because we are “finite” creatures that we can perceive spatial objects only in perceptions that exhibit the horizontal structure. Rather, it belongs to the essence of the objects themselves that they manifest themselves in this way and no other (Hua III/1: 88). According to Husserl, a perception that does not have the horizontal structure would not be a perception of a *transcendent* object, that is, an object not contained in or exhausted by my experience of it (Hua XIV: 349). A completely or “adequately” perceived tree would be no tree at all. It would rather be something inseparable from my instantaneous experience. In *Logische Untersuchungen*, Husserl takes some steps towards explaining why this is so:

If perceptions were always the actual, genuine self-presentations of objects that they pretend to be, there could be only one single perception for each object, since its peculiar essence would be exhausted in this self-presentation. (Hua XIX/2: 589)

We are to imagine that everything intended in my present tree perception is also *given* in that perception. In such a situation, if I moved my upper body just a few inches to the side, then we would have to say that my perception could no longer be a perception of the same unaltered object. This follows immediately from the assumption that *everything* intended was already given in the previous perception, i.e., that *all* aspects of the tree were fully manifest. Moving my upper body a few inches makes me see new aspects, and if these new aspects were new aspects of *the same* tree, then we would have to say that *not all aspects* were after all given to begin with, which by hypothesis they had to be (cf. Hua IX: 180-181).

But why is it important that we can have several perceptions, different in qualitative perceptual content, of the same object? If all aspects of the perceived object are fully manifest, and hence if the perceived object changes whenever the “perceptual content” changes, then the *perceived object changes with the experience*. No object could be explored further, no object could be seen from different sides, nor perceived in different ways (e.g., touched instead of watched); there would simply be nothing more to any object than what would instantaneously be given to the observer. And if this

were so, how could the experiences be experiences *of* something? Would it not rather be the case that experience and object could not be separated, that object and perceptual experience would merge? The experience would seem to absorb the object completely, so that the perceived could not be differentiated from the perceptual experience:

If we think of the ideal appearance that is extracted from the nexus of fulfilment and is temporally extended, then we would have a fully concrete appearance as the absolute givenness of a thing. What sort of givenness would this be? It would contain nothing of improper givenness, thus no back side and no interior would be without presentation. It would contain nothing indeterminate; it would be a thoroughly proper and fully determining appearance. Would there still be a difference among appearance, what appears, and the transcendence determined thereby? The appearance would indeed be no mere adumbration; it would contain no unfulfilled moments of apprehension, moments that, so to speak, point beyond themselves. (Hua XVI: 116-117)

A few pages later, Husserl draws his conclusion: “Thus the entire thing coincides, as it were, with its presentation” (Hua XVI: 120). Hence a perception *of* something, of a tree as something “out there in the garden” something transcendent in relation to the perceptual experience, is only possible in such a way that the “properly” manifest is embedded in a horizon of not-properly-manifest (cf. Hua XVI: 55). And it should be clear by now that the “core” of proper manifestation is not a potentially independent act, nor is it alone the “true” perception. The true, concrete perception is rather the *whole* that has as its ineradicable structural moments the “proper” and the “improper”.

What is the subjective motivation for this intentional surplus of “improper” manifestation?⁷ In order to see this, we must follow Husserl in considering some slightly abstract examples. First, let us disregard the surroundings in which, as explained, a perceptual object always appears. Let us then also assume that we are not yet on the level of perception of spatial objects, but rather, that we are only presented with two-dimensional visual appearances. Our task now is to account for how these

⁷ This intentional surplus, of course, can be considered a peculiar human (or mammal, or earth inhabitant) *Anschauungsform* that cannot be analyzed and explained any further. As Wittgenstein said, explanations come to an end somewhere, and perhaps we have simply reached that point here. Husserl, however, considers such postulates of inexplicable peculiarities an “asylum for phenomenological ignorance” (Hua XIII: 24).

appearances can achieve transcendent, three-dimensional significance in our perceptions.

Now, suppose I am experiencing a rectangular blue figure expanding in my visual field. There must, according to our argument above, be a possible situation where such qualitatively different “proper” perceptual content (or at least *some* qualitatively different content) functions so as to yield a perception of the same, unaltered object. One important thing here is of course the continuity of the visual experience. That is, it is essential that my experience is not like that of watching a slide show, where I see a number of discrete pictures of blue rectangular figures; rather, it is important that I experience a continuous, unbroken process (Hua XVI: 155). But this is not enough, according to Husserl. Whether there is continuity or not, the “proper” visual content, since it is changing dramatically, cannot account for the possibility of perceiving an unaltered object. This is only possible if we take something additional into account, viz. the perceiver’s immediate awareness of her position, posture, and movement (cf. Hua XVI: 160). Husserl uses the terms “kinaesthetic sensations” and “kinaesthetic sequences” to denote this subjective awareness of position and movement (Hua XVI: 161) in order to distinguish it sharply from the positions and movement of perceived objects in space.⁸ If the momentary visual content does not stand alone, as it were, but presents itself as a visual content under particular kinaesthetic circumstances out of a multitude of possible such circumstances, then it becomes intelligible that a change in visual content need not imply a change in perceived object, provided that there is an appropriate change in the kinaesthetic circumstances of the visual content. Thus, to return to our example, there is indeed such a thing as a perception of an unaltered, unmoved object, in the case where a blue rectangular figure gradually expands in the visual field; this is precisely the experience I am having at this moment, when I bend my neck and upper body in order to move closer to the *Husserliana* volume lying on my desk. My unthematic awareness of initiating a particular kinaesthetic sequence constitutes the circumstance under which such an expansion of a figure in the visual field yields a perception of an unaltered, resting book. If the same alteration in the visual field took place under different circumstances, e.g. in a situation of kinaesthetic rest, it would presumably have yielded a perception of an object moving

⁸ An additional note of caution: as the word “sensation” should indicate, Husserl is not talking about the physiological mechanisms that explain bodily movement; rather, he is talking about our subjective *awareness* of our own movement.

towards me. Likewise, we can now understand how a series of visual patterns that are even more qualitatively different (e.g., in terms of colour) can nevertheless yield perceptions of the same unaltered object. The patterns can, for example, be the visual appearances connected with a “cyclic” kinaesthetic sequence, such as a continuous movement “around” the object, seeing its different, possibly differently coloured, sides (cf. Hua XVI: 205-206).

In any visual perception of a transcendent, spatial object, then, we find a “constitutive duplicity” (Hua XI: 15). There is, on the one hand, the visual patterns, and on the other hand, the system of kinaesthetic sensations. Husserl is not only saying that both kinaesthetic sensations and visual appearances must “be there” to make a perception; he is also claiming that they must stand in a certain functional relation to each other (Hua XI: 14). The two must be correlated in such a way that *if* this particular kinaesthetic sequence is realized, *then* a particular sequence of properly given aspects follows (Hua XVI: 191; Hua IV: 57-58). The kinaesthetic sequences are the circumstances that “motivate” the visual appearances. Now we should be able to understand the horizon that makes possible the visual perception of a transcendent object. The present visual appearance is experienced as an appearance under certain circumstances, i.e. the kinaesthetic situation I have, ideally speaking at least, freely actualised.⁹ The “distribution” of proper and improper givenness I have at this moment, looking at the *Husserliana* volume, is thus the result of my having realized this possible kinaesthetic situation instead of another belonging to my open-ended system of kinaesthetic possibilities. I could choose to move closer and inspect the book from a closer range, or from other angles, whereby other aspects would be properly given (and the previously properly manifest aspects would now be empty co-intended). The horizontally intended aspects are precisely intended as *correlates* to the “kinaesthetic horizon” (cf. Hua XI: 15), the horizon of possibly realizable kinaesthetic sequences and situations. As Husserl puts it:

⁹ Obviously, I can also be kinaesthetically aware of my moving or being moved without having *chosen* to do so. If someone pushes me, I am immediately aware that my limbs change position (although I by no means chose to move them), and similarly I am, or can be, aware of my breathing. On the basis of examples such as these, Husserl in one manuscript distinguishes between “foreign” or “compulsory” kinaesthesia (someone pushes me), passive, but “allowed” kinaesthesia (my breathing – which I *could* hold back), and “active”, free kinaesthesia (I turn my head to see something) (cf. Hua XIV: 447). See the following section.

Just as the actual sequence of the *K*'s [i.e., kinaesthesia] is one out of a profusion of unitarily intimate possibilities and just as, accordingly, the *K* is at any time surrounded by a halo of apprehensional tints, i.e., *quasi*-intensions [i.e., empty intentions], which correspond to these possibilities, so the constant intention living in the appearance is encircled by a dependent halo of *quasi*-intentions which first give the appearance its determinate character, as the appearance of a thing. (Hua XVI: 190-191)

In Husserl's view, this is thus how the subject perceiving something like a tree must look: it must be a *kinaesthetic* subject, a subject in principle capable of bodily movement.¹⁰ If it were not a kinaesthetic subject, it could not perceive objects that were separate from its own experiences. This is because the object only becomes transcendent and achieves spatial significance if it shows itself perspectively. This means that there must be several qualitatively different possible perceptions of the same object. This in turn implies an "intentional surplus" reaching beyond the momentarily given visual content. Such an intentional surplus cannot be explained by way of the visual content alone. It only becomes intelligible if there is more to the perception than the visual content, if the visual patterns are appearances under particular circumstances; circumstances that have to do with the position and movement of the perceiver.

So what do Weather Watchers see? One might be tempted to reply that perhaps they cannot see objects quite as we do but at least they can see "sides" of objects. But this is clearly incoherent. It makes no sense to attribute to Strawson's Weather Watchers any kinaesthetic horizon; thus it makes no sense to attribute to them the intentional surplus that this horizon motivates. And one cannot claim that in the absence of this co-intended horizon, what I would be able to perceive would only be *sides* of things, because obviously a "side" is necessarily a side *of* an object. That is, a "side" implies other sides, thus it can only be given *as* a side when horizontal intentionality of other sides and aspects is in play (Hua XVI: 51, 55). If our Husserlian argument is correct, then Weather Watchers can perhaps experience two-dimensional

¹⁰ For more on kinaesthetic sequences and their function in perception, see Claesges (1964), Drummond (1979-80), Mulligan (1995). A good, concise introduction to the topic is found in Zahavi (2003: 98-109). Outside the phenomenological tradition, Evans has similarly insisted on the necessary relation between the subject's perceptual experience of spatial position of objects and her egocentric system of bodily capabilities. He suggests that it "does not make sense" to talk of a subject perceiving spatially distinct objects in space "unless this can be explained in terms of the subject's receiving information about the localisation of phenomena in behavioural space." That is, we can only make sense of the notion of objects having positions in space if we understand it in relation to the system of bodily orientation and movement of the perceiving subject. (Evans 1985: 396)

visual patterns, but these patterns cannot possibly achieve transcendent, spatial significance.¹¹ Maybe one can best compare the experiences possible for Weather Watchers with the visual experiences one has if one watches a television screen from almost no distance. There are coloured patterns that continuously change – but that is all. The argument outlined in this section thus constitutes a direct attack on Galen Strawson’s case for the Weather Watchers. If there is more than a merely contingent relation between perception of spatial objects and potential for bodily movement; if there is an essential connection in such a way that we can only make sense of perception of spatial locations and objects for subjects that are themselves capable of, or have been capable of, some kind of bodily movement in space, then ultimately we must conclude that Strawson’s description of the Weather Watchers does not lend itself to any intuitive fulfilment. At the very least we can say that Husserl makes such a strong case for an essential relation between kinaesthesia and visual perception that the burden of proof shifts to Strawson’s shoulders: unless he can present a full account of the perceptual constitution of spatial objects for creatures that are essentially immobile, we are entitled to dismiss his conception of the Weather Watchers as “factually absurd”.

Free kinaesthetic sequences, agency, and action

Let us at this point consider a possible critical rejoinder to our argument. We have claimed that perceptual intentionality essentially depends on the ability to realize kinaesthetic situations and sequences. In other words, we have claimed that in order for the subject to have visual perception of the world, she must be able not only to have kinaesthetic sensations, but she must also be able to freely initiate kinaesthetic sequences. But why is this last part necessary? Why would it not be enough if the

¹¹ Our claim is not that people who have been paralyzed from the neck down are unable to see things as having spatial locations, let alone that such people are unable to see spatial objects at all. We thus have no quarrel with Strawson’s Weather Watchers, given the “rooting story” Strawson mentions as a possibility. In order to help us imagine the life of Weather Watchers, Strawson introduces this story of how Weather Watchers start out as active and mobile and then, as a part of their natural course of development, gradually become rooted and immobile (Strawson 1994: 254). We do not argue that such creatures, becoming paralyzed after originally being mobile, cannot have visual perceptions. But since Strawson wants to defend the view that there is no essential link between agency and movement, on the one hand, and visual perception, on the other hand, then the rooting story, as Strawson admits, cannot be “a necessary part of the description of the Weather Watchers” anyway (Strawson 1994: 254). Our Husserlian account does, however, have certain implications for children with innate, complete paralysis. Husserl would thus be committed to the claim that children who are born completely unable to move even their eyes could not have visual perception of spatial objects in anything remotely like the usual sense.

subject just had the mere kinaesthetic sensations? This insistence on the “free initiation” seems to make us vulnerable to the following two objections: First, if “free initiation” means something like consciously choosing or deciding, then we seem to claim that in order to have perceptual experiences, we must be able constantly to deliberate how to move our bodies. Secondly, to talk about “freely initiated bodily movements” seems to imply that we are talking about actions. We would then be committed to the claim that perceptual intentionality is dependent on our ability to perform actions. From this it would seem possible to draw two conclusions that run counter to our Husserlian intuitions. First, one might conclude that perceptual and, perhaps in the final analysis, theoretical intentionality in general is dependent on practical intentionality; and while some might be attracted to such a view, it certainly is not Husserl’s. Second, one might infer that perceptual and perhaps theoretical knowledge is always governed by practical norms. That is, to perceive would be understood as if it were always in itself a practical project, and the knowledge gained from perception would always be subject to practical norms (e.g., evaluations in terms of utility and morality).¹² And surely this is absurd: my moving a few inches closer to my *Husserliana* volume while watching it is obviously not in itself, i.e. independently of any wider context, something that can be submitted to practical or moral evaluation. In order to avoid these undesirable consequences, we will have to say more about what we mean when we claim that the subject must be able to *freely initiate* kinaesthetic possibilities. First, we must attempt to establish the necessity of the *free initiation* of the kinaesthetic sequences for the perceptual experience; secondly, we must discuss this free initiation in relation to the notions of agency and action.

According to Husserl, the purely visual appearance of a thing is essentially related to two things. On the one hand, it is related to an actual sensation of the movement and position of the eye, and thereby implicitly to the whole system of possible movements and positions of the eye. On the other hand, it is also related to the actual and possible free initiation and inhibition of possible series of kinaesthetic sensations (Hua IV, §18a-b; Hua XI, §3; Hua XVI: 158, 201-202). So, if you hear a loud noise and as an immediate consequence look to the right for its source, appearances in your visual field will move to the left. In order for you to know whether it is you or the objects that move (or both), it is necessary that you have a minimal, implicit awareness of the self-

¹² We would then seem to be committed to one version of what Hurley (1998: Chap. 2) calls “the Myth of the Giving”.

initiated movement, and of multiple ways in which the free initiation of new kinaesthetic sequences will motivate changes in the appearance of the thing.

According to our argument above, the appearance of a spatial object is correlated with an awareness of one's own actual movements, and thereby implicitly also to other possible movements that are either about to be actualised, or could be, if one "chose" to. So, when the subject moves closer to an object in front of her, she does not only expect the already actualised kinaesthetic sequence to continue in a certain way, she also expects certain appearances to follow from such a continuation. And if she suddenly stops the unfolding sequence and initiates a new one – for instance, she takes a step to the right – then this will motivate a new corresponding series of appearances. Thus, as we mentioned before, the kinaesthetic sensations motivate the perceptual appearance. *If* I move in such and such a way, *then* this and that will appear. It is exactly this if-then structure of motivation that will collapse if the perceiving subject has no experience of the actual or possible free initiation. We argued above that we could never experientially go beyond two-dimensional patterns in our visual fields, were they not systematically related to the kinaesthetic circumstances in which we find ourselves. If it is true that these circumstances are essentially structured as a system of possible free kinaesthetic sequences, then it follows that proper spatial appearance is only possible for a subject that is capable of freely initiating, inhibiting, or changing a kinaesthetic sequence.

Try to imagine a subject that never had any sensation or experience of self-initiating a movement, that is, of the self-initiatedness of the movement. For this subject there would be no significant difference between, on the one hand, the experiences of initiating one's own movements, and on the other hand, the experiences of movements happening to the subject by either external force or internal reflexes. All the subject would ever have would be the mere sensation of movement. Such a subject would not be able to experience a *systematic relation* between her own movements and the appearances of things in space, nor for that matter would she be able to learn from experience about any such systematic relation. This subject would rather have the experience of an arbitrary relation between the dynamic structure of the visual field and the structure of the actual kinaesthetic sequence as it is actualised in this particular situation, but she would have no experiential grounds on which to relate it to other possible sequences and thereby to other possible appearances. This is because she would never be able, through her own controlled initiation of her bodily movements, to

associate or match this actualised relation between appearances and kinaesthetic circumstances with other relations actualised beforehand and with possible or imaginable actualisations. Kinaesthetic sequences would just be something that happened to her. There is also no reason to suppose that she could experience or ever learn the systematic relations between actual and possible movements, if she were not able to explore these relations through her own initiation. Ultimately, this would mean that no ordered system of possible kinaesthetic sequences could be constituted.

This has profound consequences. Given our argument that the horizontal structure of the object is dependent on the motivational if-then structure and thereby on the whole system of possible kinaesthetic sequences, the implication is that if there is never any experience of free initiation and therefore no such system of kinaesthetic possibilities, then the horizontal structure of the appearing object collapses. That is, the object would in this case not appear as a spatial thing in the world. To put the point in very simple terms: If I can never actively explore any perceptual object further, then why should my perceived world as such consist of objects that have perceivable back sides and inner structures? The fact that I am sometimes (passively) led to additional discoveries concerning a minor portion of this world does not really change anything, for there is no reason to think that this portion would in turn present itself as having perceivable aspects in addition to the aspects I have now passively discovered. There is no system of further appearances since there is no system of kinaesthetic possibilities for it to be correlated with.

It is necessary, then, that I am able to experience freely initiated kinaesthetic sequences, if I am to have visual perception of spatial objects. Passive or forced (“*Ichlose*”) movements alone cannot make perception possible. It is necessary that the self-moving subject is aware of her activity as her own activity, as an activity that has the I as its source, as something initiated by the I and something that can be stopped or changed by the I. In other words, it is necessary that when moving ourselves, we have a *sense of agency*.¹³ We can thus describe the kind of bodily self-awareness that corresponds to the free kinaesthetic sequences as having two structural features: 1) it involves an awareness of *what* my body is doing and 2) it involves an awareness of *who* is doing it; e.g. the awareness that my arm is moving and that I myself am moving it. We can therefore say that perceptual intentionality, as described by Husserl, necessarily involves

¹³ This concept has been introduced and applied by Gallagher in a number of recent papers, e.g. Gallagher (2000, 2004).

an experience of agency. Without kinaesthetic agency, there can be no perceptual intentionality.

But precisely what position are we adopting when we draw such a conclusion? What exactly is implied by the claim that some kind of freedom must be inherent in our experience of the kinaesthetic sequences? What precisely does it mean when we say that the subject must be able to freely initiate possible kinaesthetic sequences? What is this free initiation? Our use of words like “free” and “freedom” could perhaps suggest that we meant to say something about free will. This would mean that free initiation refers to categories such as choice, decision, practical reasoning, and last but not least, action. Our claim would then be that perceptual intentionality is dependent on the subject’s capabilities to perform freely chosen actions. However, we do not want to make such a strong claim. More specifically, we would insist that it is essential to make a distinction between agency and action, and claim that the free kinaesthetic sequences involved in perception are characterized by agency, but do not constitute actions. But in order to see why this is so, we have to provide some kind of answer to the question “What is an action?”

One minimal requirement for something to count as an action seems to be that the agent knows what she is doing.¹⁴ This is the point Davidson is apparently making in his discussion of the following example. While writing a text, I intend to produce ten carbon copies (Davidson 1971: 50). This action is of course not dependent on my knowledge of whether or not I actually succeed in producing the ten copies; rather, the knowledge is delimited to my first-person perspective in action. That is to say, my knowledge is delimited to an awareness of my intention (to produce ten carbon copies) and to my actual bodily performance. So, for the behaviour to count as an intentional action the agent must have some knowledge of what she is trying to do and of what she is actually doing with her body; that is, her bodily movements must be intentional under a description (*ibid.*: 51, 54; see also Davidson 1987: 39, on the monitoring function of intentions). It seems fair to say that if the agent did not know that she was moving her hand, or if she would deny that it was she herself who moved it, then we would not say that she was writing and *a fortiori* that she was trying to produce ten carbon copies.

¹⁴ For an early recognition of the relation between intention and first-person related knowledge, see Anscombe (1957 [2000]: §§8: 28-29). A recent volume edited by Roessler & Eilan (2003) testifies to a renewed interest in the problem.

The agent of an action must therefore have a twofold knowledge, namely knowledge of her intention and knowledge of her own bodily acts.

It seems clear that we must be careful about what we understand by “knowledge”, however. First of all, we cannot be dealing with thematic, observational knowledge here. Rather, the “knowledge” in question must be allowed to be of a non-observational and non-inferential character, an implicit form of bodily self-awareness. Further, if “knowledge” is understood as implying the ability to pass judgement and to give reasons, then we will end up with too limited a concept of action which excludes all those actions that we are unable to give rational reasons for.¹⁵ Yet we take it as a fact that people do experience unpremeditated and absentminded actions like putting the keys in your pockets after closing the door, as actions – even though you did not have any explicit reasons for performing the action prior to your initiation of it. Thus, we must also allow the notion of action to be divorced from the notion of conscious decision. We often perform actions that seem to be just as non-deliberate and “automatic” as the freely initiated kinaesthetic sequences in perception. For instance, while engaged in a discussion, we can take a drink from a glass of water or adjust the glasses on our noses without noticing it. In this case, we do not perform any deliberation or make any conscious choice or decision about what to do. We just do it. But it would nevertheless be very counterintuitive to claim that the example does not involve actions.

Yet all this seems to match our description of the bodily sense of agency perfectly. We almost never initiate a kinaesthetic sequence because we consciously decide to do so. There is no decision or practical deliberation involved in the experience of free initiation that we have been discussing. If I hear a loud crash to the right of me and turn my head to look, then I just do it. There are no intermediary conscious states such as decisions or intentions; nor do I have any thematic awareness of turning my head. But precisely the same must be allowed for actions, as we have just seen. Thus, if there is a difference between the freely initiated kinaesthetic sequences in perception, and action proper, it relies on neither conscious decision, nor thematic knowledge, nor the ability to provide rational reasons or justifications.

In order to see what might constitute such an essential difference between the experience of agency and action, let us look at a concrete example. Imagine that you

¹⁵ This seems to be Velleman’s position (2000: 2, 26).

discover that the shoelaces on your left shoe are untied. You see the untied shoelaces and desire that they be tied, but you are in the middle of a busy street and focus on crossing it unscathed, for which reason you decide to tie the laces once you have arrived safely on the other side. In other words, you form what has been termed a future-directed intention. Having crossed the street you remember your prior intention to tie the laces and proceed to do so immediately.¹⁶ What happens here? The prior intention makes us settle on a course of action and makes us ready to initiate it when the time comes, but it is not responsible for the initiation. The agent still needs to get herself going, a starting push, an “*anfänglichen Willensimpuls*”, as Melle (1992: 292) terms it. As if one were saying, “Ok, *now* I tie my laces”, only that this *conative push* is not of verbal or even quasi-verbal nature. It is not a “mental” experience nor a “physical” one, like a tactile sensation or a feeling of pain, but something in between, however one is to describe that. It is essentially an experience of transition, of going from a state of inactivity to action or from one action to another. The conative push or *fiat*, as Husserl calls it with reference to James, is the starting off of the activity that constitutes the action (Hua XXVIII: 109-112). We do not mean to say that the agent has an experience of a conative push, on the one hand, and the beginning of the bodily activity of tying the shoelaces on the other. The conative push is exactly the experience of starting off the activity. But it is not only that. The conative push is an intentional experience directly related to the intentional goal of the whole activity, in our example “having tied the laces”, and only secondarily related to the first movement of the ensuing activity.¹⁷ As such, the conative push is an experience that is not fulfilled by the fact that the agent starts to move; rather, it is fulfilled by the completion of the whole activity leading to the goal. In other words, the conative push is intentionally related to the completion of the bodily activity, the initiation of which is the immediate content of the experience.

¹⁶ For a further discussion of the re-presentation of prior intentions and their transformation into intentions in action, see Hua XXVIII: 109, and Ginet (1990: 142-144). For the concepts of “prior intention” and “intention in action” and for a causal account of their mutual relation, see (Searle 1983: Chap. 3).

¹⁷ A. Mele (1992) has a similar analysis of action as constituted by a proximal intention and ensuing bodily movements guided by the intention, but in his analysis they are exactly related as cause and effect, which seems to imply that they could in principle be given in isolation of each other. This separation between the mental cause and the behavioural effect is especially clear in the so-called new volitionist theory of action. Contrary to Husserl, most volitionists insist on the possibility of isolating the mental aspect of action as a single mental act, often called a “trying”. It is important to note that on our account the “conative push” cannot be actualised as a separate mental act; it is nothing more than the experience of getting a bodily activity started that has a certain expected practical goal.

But does this account exclude unpremeditated, absentminded actions, such as making sure the door is locked before leaving, scratching one's arm because it itches etc.? In such cases, we have no conscious experience of a conative push that initiates the activity, but we nevertheless experience performing actions.

When I absentmindedly scratch my arm, there is no *patent* consciousness of a "push", that is, of starting off the activity. But could there be a *latent* form? The conative push is not only an experience of getting started. It is not to be understood as a push in terms of ballistics, as when the force of one entity sets another one in motion. The conative push is not a causal force but an intentional experience, which has the completion of the bodily activity as its object. It is an experience of "getting started with...", which has an intentional goal that gives the whole activity its unity. The intentionality of the conative push unites our flowing bodily activities into unities in which the bodily activities are given a functional role. The grasping of a glass can be part of cooking in one situation and of taking a drink in another, and in yet another circumstance part of a murder, depending on the intentional goal of the action. Thus, the experience of wilfully initiating a bodily activity is related to this bodily activity as that which gives it structure and meaning, and the bodily activity is related to the conative experience as a carrying out, as a fulfilling activity. In other words, the conative experience is not only the experience of getting the fulfilling activity started, but also of intending the completion or fulfilment of the activity as a whole. This structure is the reason why we can have the experience of finding ourselves in the middle of an *action*; because even though we had no patent experience of a conative push, we still experience our activities as directed toward some goal or completion, and this is enough.

Let us summarize what we, following Husserl, take to be necessary traits of our first-person experience of action. Firstly, there is the patent or latent conative experience of getting the activity going and directing it towards an intended completion. Secondly, there is the bodily activity that, phenomenologically speaking, is constituted as a continuous projection or protention of the next-coming phases of the activity as well as the continuous fulfilment of the projection of the preceding phase.¹⁸ And through all these phases goes the intentional "beam" of the conative experience uniting

¹⁸ This is by no means an exhaustive or satisfying description. For one thing we have left out the constitutive "retentional" aspect of the experience. See Hua XXVIII: 110. For more on Husserl's phenomenology of will and the involved fulfilment structures, see Melle (1992).

the phases and giving them a functional role in regards to the intended completion. Thus, the intentional bodily activity has a complex intentional structure involving both “empty” intentions (the protentions or expectations) and fulfilling intentions, in this case kinaesthetic sensations and perception (visual, tactile, etc.). So, when you embark on the activity of walking to the kitchen, not only do you expect in each phase a certain sequence of kinaesthetic sensations but also perceptual changes, e.g. things moving in certain expected directions in your field of vision. It is part of our experience of performing a willed activity and of being in control that these expectations are continuously fulfilled.¹⁹ From this description it follows that the experience of performing a willed activity involves the experience of agency but is not identical to it. The experience of agency was described above as an awareness of the activity or movement as initiated and possibly inhibited by the I. But the experience of will in action involves more than this; it involves being directed towards the completion of the activity in question. An intended practical completion that does not lie at the end of each single movement, but on the contrary unites the movements, makes them into phases of a whole in which they are given a certain internal structure and role. The intentional bodily activity only has its meaning as an action or parts of an action through the overarching intentionality of the conative experience.

Finally, it has become clear that we should not simply regard agency as the first-person experiential counterpart of action, even though the experience of agency is a necessary aspect of the experience of action. More formally put, we can say that the experience of action entails the experience of agency, whereas the experience of agency does not entail an experience of action. We can have an experience of agency without performing an action. The free initiation of the subject’s possible kinaesthetic sequences, i.e. the experience of agency, is an aspect of both perception and action. However, movement (or rather, moving) has a different experiential function and status for the subject depending on the intentional frame in which it is contained. In perception it primarily has the function of motivating appearances; in action the activity is intentionally related not only to the next-coming phase of the activity, but also through the intentionality of the conative experience (the “push”) to the practical

¹⁹ This has also recently been argued by Roessler (2003). For a different opinion about the phenomenology of action and will, see Smith (1992). Compared to our analysis, he assigns a merely secondary role to the kinaesthetic and perceptual features of the experience.

completion and thereby also to the either patent or latent experience of the conative push itself.

Concluding this section, we can say that perceptual intentionality by virtue of its dependence on free kinaesthetic sensations necessarily involves the experience of agency. As we understand the concepts, it makes sense to say that we cannot have perceptual intentionality without agency. However, we can have perception without action. Action is therefore only instrumentally related to perception and not constitutively, whereas perception is constitutive to the experience of action, given the picture of the involved fulfilment structures we have loosely sketched.

Conclusion

If our findings in the previous three sections are correct, then the following points have been established. First, one should not insist on Hume's division of human knowledge. There are essential insights, for which the contrary claims are neither logically incoherent, nor of course contingently false, but rather "factually absurd". Second, Strawson's description of the Weather Watchers belongs to precisely this category: when we probe into the conditions for the possibility of visual perception, we realize that we can only make sense of visual perception for subjects able (at least in principle) to move themselves, i.e. subjects able to engage in what Husserl calls free kinaesthetic sequences. This means that Strawson's Weather Watchers will at most be able to perceive two-dimensional coloured patterns; they would not be able to have anything resembling our perceptions of spatial objects. Third, the free kinaesthetic sequences must be understood as involving a form of agency. This means that visual perception essentially presupposes agency. However, contrary to the views of some phenomenologists, we should hesitate to draw the further conclusion that the capacity for visual perception presupposes actions. While it makes perfect sense to say that kinaesthesia involves agency, there is no justification for the claim that it involves action.²⁰

²⁰ Thor Grünbaum's work on this article was funded by the Danish National Research Foundation. Soren Overgaard's work was funded by the Carlsberg Foundation. The study was carried out at the Danish National Research Foundation: *Center for Subjectivity Research*. We are grateful to Lisa Käll, Dan Zahavi, and the editors of *Cognitive Semiotics* for comments on earlier drafts of the paper.

References

Works by Husserl:

- Husserliana*, Vol. III/1. *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie. Erstes Buch: Allgemeine Einführung in die reine Phänomenologie*. Edited by Karl Schuhmann. The Hague: Martinus Nijhoff, 1976.
- Husserliana*, Vol. IV. *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie. Zweites Buch: Phänomenologische Untersuchungen zur Konstitution*. Edited by Marly Biemel. The Hague: Martinus Nijhoff, 1952.
- Husserliana*, Vol. VI. *Die Krisis der europäischen Wissenschaften und die transzendente Phänomenologie: Eine Einleitung in die phänomenologische Philosophie*. 2nd edition. Edited by Walter Biemel. The Hague: Martinus Nijhoff, 1976. *The Crisis of European Sciences and Transcendental Phenomenology*. Translated by David Carr. Evanston, Ill.: Northwestern University Press, 1970.
- Husserliana*, Vol. IX. *Phänomenologische Psychologie: Vorlesungen Sommersemester 1925*. Edited by Walter Biemel. The Hague: Martinus Nijhoff, 1962.
- Husserliana*, Vol. XI. *Analysen zur passiven Synthesis: Aus Vorlesungs- und Forschungsmanuskripten 1918-1926*. Edited by Margot Fleischer. The Hague: Martinus Nijhoff, 1966.
- Husserliana*, Vol. XIII. *Zur Phänomenologie der Intersubjektivität: Texte aus dem Nachlass. Erster Teil: 1905-1920*. Edited by Iso Kern. The Hague: Martinus Nijhoff, 1973.
- Husserliana*, Vol. XIV. *Zur Phänomenologie der Intersubjektivität: Texte aus dem Nachlass. Zweiter Teil: 1921-1928*. Edited by Iso Kern. The Hague: Martinus Nijhoff, 1973.
- Husserliana*, Vol. XVI. *Ding und Raum: Vorlesungen 1907*. Edited by Ulrich Claesges. The Hague: Martinus Nijhoff, 1973. *Thing and Space: Lectures of 1907*. Translated by Richard Rojcewicz. Dordrecht: Kluwer Academic Publishers, 1997.
- Husserliana*, Vol. XIX/2. *Logische Untersuchungen. Zweiter Band: Untersuchungen zur Phänomenologie und Theorie der Erkenntnis. Zweiter Teil*. Edited by Ursula Panzer. The Hague: Martinus Nijhoff, 1984. *Logical Investigations. Volume Two*. Translated by J. N. Findlay. London: Routledge and Kegan Paul, 1970.
- Husserliana*, Vol. XXVIII. *Vorlesungen über Ethik und Wertlehre 1908-1914*. Edited by Ullrich Melle. Dordrecht: Kluwer Academic Publishers, 1988.

Works by Other Authors:

- Anscombe, G.E.M. (2000). *Intention*. Cambridge, Mass.: Harvard University Press.
- Claesges, U. (1964). *Edmund Husserls Theorie der Raumkonstitution*. The Hague: Martinus Nijhoff.
- Danto, A.C. (1973). *Analytical Philosophy of Action*. Cambridge: Cambridge University Press.
- Davidson, D. (1971). 'Agency'; in *Essays on Actions and Events*. 2nd edition, Oxford: Clarendon Press, 2001.
- Davidson, D. (1987). 'Problems in the Explanation of Action'; in P. Petit, R. Sylvan & J. Norman (eds.), *Metaphysics and Morality*. Oxford: Basil Blackwell.
- Dennett, D. (1982). 'Where am I?'; in D. R. Hofstadter & D. C. Dennett (eds.), *The Mind's I: Fantasies and Reflections on Self and Soul*: 217-231. Toronto: Bantam Books.
- Drummond, J. J. (1979-80). 'On Seeing a Material Thing in Space: The Role of Kinaesthesia in Visual Perception'. *Philosophy and Phenomenological Research*, 40: 19-32.
- Evans, G. (1985). 'Molyneux's Question'; in *Collected Papers*. Oxford: Clarendon Press.
- Gallagher, S. (2000). 'Philosophical conceptions of the self: implications for cognitive science'. *Trends in Cognitive Sciences*, Vol. 4, No. 1: 14-21.
- Gallagher, S. (2004). 'Agency, ownership, and alien control in schizophrenia'; in D. Zahavi, T. Grünbaum & J. Parnas (eds.), *Interdisciplinary Perspectives on Self-Consciousness*. Amsterdam: John Benjamins Publishing Company.
- Ginet, C. (1990). *On Action*. Cambridge: Cambridge University Press.
- Hume, D. (1993). *An Enquiry Concerning Human Understanding*, ed. E. Steinberg. Indianapolis: Hackett Publishing Company.
- Mele, A. (1992). *Springs of Action*. Oxford: Oxford University Press.
- Melle, U. (1992). 'Husserls Phänomenologie des Willens', *Tijdschrift voor filosofie*, 54: 280-304.

- Mulligan, K. (1995). 'Perception'; in B. Smith & D. W. Smith (eds.), *The Cambridge Companion to Husserl*. Cambridge: Cambridge University Press.
- Parnas, J. and Zahavi, D. (2000). 'The Link: Philosophy – Psychopathology – Phenomenology'; in D. Zahavi (ed.), *Exploring the Self*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Roessler, J. & Eilan, N. (eds.) (2003). *Agency and self-awareness*. Oxford: Clarendon Press.
- Roessler, J. (2003). 'Intentional Action and Self-Awareness'; in J. Roessler & N. Eilan (eds.), *Agency and self-awareness*. Oxford: Clarendon Press.
- Sartre, J.-P. (1943). *L'être et le néant*. Paris: Éditions Gallimard.
- Searle, J. (1983). *Intentionality*. Cambridge: Cambridge University Press.
- Sheets-Johnstone, M. (1999). *The Primacy of Movement*. Amsterdam: John Benjamins Publishing Company.
- Smith, D.W. (1992). 'Consciousness in Action'. *Synthese*, 90: 119-143.
- Strawson, G. (1994). *Mental Reality*. Cambridge, Mass.: The MIT Press.
- Velleman, J.D. (2000). *The Possibility of Practical Reason*. Oxford: Clarendon Press.
- Zahavi, D. (1999). *Self-Awareness and Alterity*. Evanston: Northwestern University Press.
- Zahavi, D. (2003). *Husserl's Phenomenology*. Stanford: Stanford University Press.

Paper received April 2004; revised February 2007

Shaun Gallagher

Sense of Agency and Higher-Order Cognition: Levels of explanation for schizophrenia¹

I contrast two general approaches to the explanation of certain symptoms of schizophrenia: a top-down model that emphasizes the role of introspection, and a bottom-up model that takes both neurological and phenomenological aspects into account. Top-down accounts look to higher-order cognitive attributions as disrupting a sense of self-agency; bottom-up accounts suggest that the phenomenal sense of agency is generated by specific neuronal mechanisms, and that these mechanisms are disrupted in cases of schizophrenia. The bottom-up view is supported by both clinical reports and experimental science, and is consistent with a phenomenological approach to psychopathology.

Introduction

Philosophers and cognitive neuroscientists are interested in psychopathologies primarily for theoretical reasons. They often appeal to what goes wrong in pathological cognition to illuminate how things work in “normal” cognition. While psychiatrists are primarily concerned with treatment and therapy, theoreticians can learn from their clinical accounts and especially from the patient’s own narratives about their experience. Phenomenological approaches are characterized as starting with experience rather than with pre-established theories. Phenomenological psychiatrists and philosophers take the patient’s first-person narratives seriously. That is, they regard them as reflective of the

¹ An earlier version of this paper was presented at the European Society for Philosophy and Psychology, Lyon, France (July 2002)

patient's actual experiences. Experience, itself, however, is complex. One can distinguish between (1) first-order phenomenal experience, that is, the immediately, pre-reflective, lived-through experience of the world, and (2) higher-order cognition, a reflective experience which supports the ability to make attributive judgments about one's own first-order experience. Cognitive neuroscientists are interested in explaining how such experiences are generated in a third level of the cognitive system, (3) the non-conscious, sub-personal processes that are best described as neuronal or brain processes.²

According to what Francisco Varela (1996) has called *neuropsychophenomenology*, cognitive neuroscience ought to be guided in some way by the experience that it attempts to explain. A neuroscientist who, for example, is using neuroimaging techniques, will look for very different things if she starts with the idea that schizophrenic symptoms like thought insertion and delusions of control are caused by a dysfunction of higher-order cognition than if she starts with the idea that schizophrenic symptoms are primarily manifested at the level of first-order experience. These two starting points represent a contrast between a *top-down* explanation and a *bottom-up* explanation of schizophrenic symptoms. In this paper I want to make this distinction clear, and argue that to explain positive symptoms of schizophrenia, like thought insertion and delusions of control, within the phenomenological-neurological framework, one needs to take a bottom-up approach, and that involves attending to the first-person experience of patients.³

Distinctions between self-agency and self-ownership in schizophrenia

Distinctions between self-agency and self-ownership may be found both in first-order phenomenal experience and higher-order consciousness. In regard to the latter, for example, Graham and Stephens (1994) work out their account of introspective alienation in schizophrenic symptoms of thought insertion and delusions of control in terms of two kinds of self-attribution.

² There may be some intermediate level(s) of description (syntactical or representational) that are also understood as non-conscious, but for purposes of this paper I will leave this possibility aside. Also, I don't deny that there may be unconscious mental states, but I will also leave this idea aside. Assume also that experience is not simply cognitive, but also emotional and embodied.

³ Similar distinctions serve as a framework for the analysis found in Gerrans (2001), and they are consistent with his remarks there, and with Gallagher (2000a and 2000b, 2004). The approach taken here fits into the general contours of a phenomenological approach to schizophrenia, the history and main points of which are outlined by Aaron Mishara (National Institute of Mental Health) in an unpublished text, "Agency and Self in Schizophrenia: Should Icarus More Fear the Heights or the Depths?" Mishara distinguishes between an Apollonian (top-down) and Dionysian (bottom-up) approach. For a good example of a clinically informed, phenomenological, bottom-up approach, see Parnas (2003) and Sass (2000, 1998).

—**Attributions of subjectivity (ownership):** the subject reflectively realizes and is able to report that he is moving or thinking. For example, he can say, “This is my body that is moving.”

—**Attributions of agency:** the subject reflectively realizes and is able to report that he is the cause of his movement or thinking. For example, he can say “I am causing this action.”

This distinction seems consistent with another very similar distinction made at the level of first-order phenomenal consciousness.

—**Sense of ownership:** the pre-reflective experience or sense that I am the subject of the thought or movement (“I am thinking,” or “I am moving”).

—**Sense of agency:** the pre-reflective experience or sense that I am the cause or author of the thought or movement.

These two distinctions are not identical, even if they are consistent or similar.⁴ They differ in regard to the level or order of self-conscious experience. The distinction between sense of ownership and sense of agency belongs to a first-order phenomenology; it involves a pre-reflective (non-conceptual) sense of ownership and agency implicit to experience, and generated by neurological processes responsible for personal control of action or thought.⁵ In contrast, when one introspectively *attributes* ownership (subjectivity) and agency, one takes a second- or higher-order attitude

⁴The distinction understood in this way is made in Gallagher (2000a, 2000b). I am suggesting that this is not the same distinction found in Graham and Stephens (1994), although in their later work (Stephens & Graham 2000) they do make a very similar one. As I will make clear in the following, however, Stephens and Graham explain the distinction between a sense of agency and a sense of ownership as a production of higher-order cognition that is read into the level of phenomenal consciousness. Thus, what Stephens and Graham characterize as a sense of agency that is “a normal component or strand in our experience of thinking,” but “normally phenomenologically intertwined with introspective awareness as well as with the sense of subjectivity” (2000: 9) turns out, on their view, to be “constituted by our self-referential narratives or conceptions of our underlying intentional states” (2000: 183). The difference in explanation (and perhaps in phenomenology) between a sense of agency generated “from below” by neurological processes, and a sense of agency generated “from above” by cognitive processes goes to the heart of the matter as it is explicated in this paper.

⁵ Recent studies based on brain imaging suggest the involvement of the right inferior parietal cortex, the anterior insula, and other areas (Chaminade & Decety 2002, Farrer & Frith 2001) in the generation of the sense of agency. See below.

toward first-order phenomenal experience (see Lambie & Marcel 2002, Gallagher & Marcel 1999). Thus, in explanations of schizophrenic symptoms such as delusions of control, thought insertion, and auditory hallucination, these distinctions work at different levels. More generally they help to distinguish between cognitive (higher-order, top-down) versus phenomenological (first-order, bottom-up) accounts of schizophrenia.

The distinctions between ownership and agency at either level can be worked out by considering the difference between voluntary and involuntary movement. If, for example, someone pushes me from behind, I sense that it is my body that is moving (ownership/subjectivity), but that I did not cause the movement (no agency). This is like the complaint in schizophrenic delusions of control: My body moved (ownership), but I did not cause the movement (no agency); moreover, the subject complains, someone else caused the movement. Frith provides the following example of a patient who attributes his own movement to an alien agency.

‘The force moved my lips. I began to speak. The words were made for me’.
(Frith 1992: 66).

A similar example is given by Mellor.

‘When I reach my hand for the comb it is my hand and arm which move, and my fingers pick up the pen, but I don’t control them’. (Mellor 1970: 17)

One can also understand thought insertion in these terms. In the ordinary case of involuntary or “unbidden” thoughts (Frankfurt 1976) I might say that there is a thought in my stream of consciousness, that is, it is a thought that I am experiencing (ownership), but that I did not intend to think this thought (no agency), although in such cases, of course, I can certainly attribute agency to myself. That is, I know at a second-order level that this unbidden thought was generated by me. In contrast, however, the schizophrenic would deny his own agency and may attribute the cause to someone or something else.

‘I look out my window and I think that the garden looks nice and the grass looks cool, but the thoughts of Eamonn Andrews come into my mind. There are no

thoughts there, only his He treats my mind like a screen and flashes his thoughts onto it like you flash a picture'. (Mellor 1970: 17).

In all of these cases, there is a lack of a sense of agency for the action or thought, and for the schizophrenic the problem seems to be about agency rather than about ownership. Schizophrenics who suffer from these symptoms acknowledge that they are the ones that are moving, that the movements are happening to their own body, or that thoughts are happening in their own stream of consciousness, but they claim they are not the agents of these movements or thoughts – when in fact they do cause the movement or thought.

Problems with top-down accounts

I want to suggest that accounts of problems with self-agency are best developed from the bottom up – specifically, that a neurological account helps to explain a failure manifested in first-order experience, and that just such a failure helps to explain why there is a misattribution of agency at the higher-order level. On this approach, the majority of the explanation needs to be worked out at the neurological and phenomenological levels, where the latter involves an account of what it's like for the subject – an analysis of the basic first-order phenomenal experience involved in the *senses* of ownership and agency or their disruption. Once these analyses are in place, what happens at the cognitive level of self-attribution (the *attribution* of ownership but not of agency, and the misattribution of agency to another) is straight-forwardly explained as an effect of the lower-order disruptions.

This bottom-up strategy is to be distinguished from a top-down account that bestows causal power in this regard to the higher-order cognitive or attributive level. On the latter approach, the misattribution of agency becomes the *explanans* rather than the *explanandum*. Graham and Stephens, (1994, and Stephens & Graham 2000), for example, attempt to work out this type of top-down account. Following Dennett and Flanagan, they propose an explanation of the sense of agency in terms of “our proclivity for constructing self-referential narratives” which allow us to explain our behavior retrospectively: “such explanations amount to a sort of theory of the person’s agency or intentional psychology” (1994: 101, Stephens & Graham 2000: 161).

[W]hether I take myself to be the agent of a mental episode depends upon whether I take the occurrence of this episode to be explicable in terms of my underlying intentional states (1994: 93).

This top-down account depends on a “theory of mind” approach according to which we reflectively make sense of our actions in terms of our beliefs and desires. So, if a patient does or thinks something for which he has no intentions, beliefs, or desires – mental states that would normally explain or rationalize such actions – the first-order movements or thoughts would not appear as something he intentionally does or thinks. Thus, whether something is to count for me as my action

depends upon whether I take myself to have beliefs and desires of the sort that would rationalize its occurrence in me. If my theory of myself ascribes to me the relevant intentional states, I unproblematically regard this episode as my action. If not, then I must either revise my picture of my intentional states or refuse to acknowledge the episode as my doing. (1994: 102).

On this approach, non-schizophrenic first-order phenomenal experience appears the way it does because of properly ordered second-order interpretations, and schizophrenic first-order experience appears the way it does because of second-order *misinterpretation*.

[T]he subject’s sense of agency regarding her thoughts likewise depends on her belief that these mental episodes are expressions of her intentional states. That is, whether the subject regards an episode of thinking occurring in her psychological history as something she does, as her mental action, depends on whether she finds its occurrence explicable in terms of her theory or story of her own underlying intentional states. (Graham & Stephens 1994: 102; see Stephens & Graham 2000: 162ff).

It would follow from this view that schizophrenic symptoms are inferential mistakes made on the basis of higher-order introspective or perceptual self-observations. On the account offered by Graham and Stephens, “what is critical is that the subject finds her thoughts inexplicable in terms of beliefs about her intentional states” (1994: 105). On this theory of mind approach, subjects normally attempt to explain their own

experience to themselves. Graham and Stephens suggest that (schizophrenic) mistakes in such explanations may be motivated by negative emotions concerning the thoughts or movements in question, but that they need not be so motivated. They point out that this is complicated by the fact that in many cases, the thoughts/movements are innocuous or neutral. In such cases no evaluative or emotive aspects need be involved. In effect, the failure to attribute agency or the misattribution of agency may simply be a (mis)judgment made about the incongruence between the content of the current experience and what the subject takes to be her more general conception of her intentional states (see Stephens & Graham 2000: 170).

Graham and Stephens's top-down explanation ignores first-level phenomenology and has nothing to say about neurological processes that may be involved. In line with higher-order-representational theories of consciousness, for Graham and Stephens the first-order experiences of schizophrenic symptoms are not lived through in any originally conscious sense, but are determined by theoretical mistakes made at higher cognitive levels. Thought X seems not to be my thought only *after* some reflective verification process has failed.

This is not an uncommon type of analysis. Ralph Hoffman (1986) proposes an explanation of verbal hallucinations in schizophrenia that depends on the failure of a self-corrective, introspective judgment that normally corrects the default and automatic inference that ordinary unintended instances of inner speech are caused by something other than myself. On this view, a default mechanism defines non-intentional actions/thoughts as not caused by me – and in schizophrenia this default continues to function. In the non-schizophrenic case, however, on some second-order level, there is a properly functioning “self-corrective process” that vetoes this mechanism and verifies that in fact such actions are my own. One learns that “unintended or alien representations occur during prior passive states and thereby dismiss their veracity” (Hoffman 1986: 509). In schizophrenia this second-order cognitive correction fails and leads to the misattribution.

Stephens and Graham (2000) contest Hoffman's view because it fails to explain why we correctly attribute agency for non-self-generated speech to others when we are in a passive (listening) state and the speech is in fact generated by others. Hoffman's self-corrective process, as he describes it, would prevent us from doing so. But what Stephens and Graham offer as “a more plausible version” is more plausible only if one

accepts the idea that there is something like a second-order cognitive deliberation that occurs each time we hear someone else speak or generate our own inner speech. They suggest that the self-corrective process should be thought of as a judgment-withholding process which “induces the subject to withhold judgment on or reconsider such nonself inferences” (2000: 107). This second-order cognitive process would have to consider a variety of evidence about whether it is one’s own (internal) or someone else’s (external) speech.

Although Stephens and Graham do not provide any indications of what may be happening in the brain that would cause the introspective problems in schizophrenia, there are some top-down explanations that are interestingly combined with neurological explanation. For example, Frith’s (1992) account of inserted thought appeals to problems on the metarepresentational level, where metarepresentation is defined as a full-fledged second-order act of reflection. The failure of metarepresentational introspection is attributed to neurological dysfunctions associated with sub-personal efferent copy at a brain mechanism called the comparator. It can be argued, however, that Frith’s account misconstrues the phenomenological (first-order) level of experience, mistakenly correlates brain mechanisms responsible for first-order experience to second-order cognition, and fails to explain the *mis*attribution of agency (Gallagher 2000, Gallagher 2004b; also see Stephens & Graham 2000: 141ff).⁶

John Campbell (1999), although claiming to follow Frith, moves, I think, in a slightly different direction. For him, the problem is not with the introspective or metarepresentational level of self-monitoring. Second-order (introspective) processes play no causal role here, and in fact, on his view, second-order self-monitoring functions must continue to work properly even for cases of inserted thought. Campbell, however, like Frith, mixes second-order cognitive processes and neurological processes that correlate with first-order experience⁷ and yet does not consider the possibility that first-order experience might in fact play a role in the problem.

⁶ Frith now acknowledges the problems in his 1992 account for explaining thought insertion, although not necessarily for delusions of motor control (see Frith 2004).

⁷Campbell (1999) writes: “It is the match between the thought detected by *introspection*, and the content of the efferent copy picked up by the comparator, that is responsible for the sense of [agency for] the thought. ... You have knowledge of the content of the thought only through introspection. The content of the efferent copy is not itself conscious. But it is match at the monitor [= comparator] between the thought of which you have introspective knowledge and the efferent copy that is responsible for the sense of being the agent of that thought. It is a disturbance in that mechanism that is responsible for the schizophrenic finding that he is introspectively aware of a thought without having the sense of being the agent of that thought.”

A bottom-up account

A good example of a phenomenologically-informed bottom-up approach is given by Louis Sass (1992, 1998). His account is quite in contrast to Stephens and Graham.⁸ On Sass's account, neurological problems may cause tacit sensory-motor processes that are normally implicit in first-order phenomenal experience to become abnormally explicit. This becoming explicit is already a form of automatic, or what Sass (2000) calls "operative" hyperreflexivity, and it may motivate more willful forms of hyper-reflective awareness (also see Sass 2003, Sass & Parnas 2003). The normally tacit integration of cognitive, emotional, and motivational experience is disrupted in schizophrenic experience at the phenomenal level; the implicit unity of the self breaks down; and one begins to feel alienated from one's thoughts and actions. For Sass, this more primary disruption often brings on reflective forms of hyperreflexivity involving an excessive type of second-order introspection. Though secondary and defensive in a causal sense, this introspective hyper-reflection often plays an important causal role in bringing on problems of its own. To the extent that higher order processes do play a causal role in bringing about self-alienation, this occurs not as the original source of the problem, but as part of a cascade of processes that begin on a lower level.

Sass thus presents a view of the potentially complex interactions that can occur between more automatic, lower-level processes and higher-level ones that may have a more willful and possibly defensive quality. The account that I outline in this section pursues this kind of bottom-up perspective. It takes the distinction between sense of ownership and sense of agency at the level of first-order phenomenal experience seriously, and presents a bottom-up model consistent with recent empirical findings.

On this bottom-up account, problems with self-agency that manifest themselves in thought insertion and delusions of control are generated on a neurological level. The neurological picture is complex, but recent results of brain-imaging studies suggest the importance of two neuronal areas in generating a sense of agency for movement.

Farrer and Frith (2001) have shown contrasting activation in the right inferior parietal cortex for perception of action caused by others, and in the anterior insula

⁸Stephens and Graham (2000) misinterpret Sass's work in this regard. They view Sass as proposing a top-down explanation, bestowing causal power on hyper-reflective introspection. But Sass (1992) explicitly claims that a higher-order introspection need not be the origin of hyperreflexivity or self-alienation, and indeed, that lower-level processes are probably prior. He attributes a possible causal role to the disruption or under-activation of more automatic and less volitional neurophysiological processes (1992: 69, 386). For Sass's brief critique of Graham and Stephens, see Sass (1999: 261-62).

bilaterally when action is experienced as caused by oneself. One possible explanation for the involvement of the right inferior parietal cortex in the discrimination between self-agency and other-agency is suggested by Jeannerod (1999). Namely, actions performed by others are perceptually mapped in allocentric coordinates. Farrer and Frith note that “there is strong physiological evidence that the inferior parietal cortex [involves this] kind of remapping ... to generate representations of body movements in allocentric coordinates” (2001: 601).

In contrast to the function of the right inferior parietal cortex, Farrer and Frith suggest that the role of the anterior insula in providing a sense of self-agency involves the integration of three kinds of signals generated in self-movement: somatosensory signals (sensory feedback from bodily movement, e.g. proprioception), visual and auditory signals, and corollary discharge associated with motor commands that control movement. “A close correspondence between all these signals helps to give us a sense of agency” (2001: 602).⁹

Other studies show that lesions in the right parietal lobe can lead to difficulties in attributing actions. For example, lesions in the right parietal cortex can cause a disturbance of the sense of ownership for one’s limbs (as in neglect or alien hand syndrome). Also, in psychiatric and neurological patients self-awareness disorders have been linked to metabolic abnormalities in the right inferior parietal cortex. In schizophrenic patients the feeling of alien control (delusions of control) during a movement task has been associated with an increased activity in the right inferior parietal lobe (Spence *et al.*, 1997).

Of course things are likely more complicated than this in both normal and schizophrenic experience. The sense of agency for motor action may depend on a pre-action, forward motor control mechanism that matches motor intention and efference copy of the motor command. The proper timing of such a match may depend on the proper functioning of the supplementary motor area, the premotor, and prefrontal cortexes, and such functions are known to be disrupted in schizophrenic subjects with delusions of control (Fournieret & Jeannerod 1998, Georgieff & Jeannerod 1998; see

⁹ Studies by Decety *et al.* (2002), Chaminade and Decety (2002) and Farrer, Franck, Georgieff, Frith, Decety, and Jeannerod (2003) support this conclusion. It is important to note that in the experiments mentioned here the authors have adopted the same concept I have defined above as the sense of agency, and as outlined in Gallagher (2000a). Other empirical studies consistent with the findings mentioned here have also used this definition (see, e.g., Blakemore *et al.* 2000, Fournieret *et al.* 2001, Jeannerod 2003, Ruby & Decety 2001, van den Bos & Jeannerod 2002, Vogeley *et al.* 2001, Vogeley & Fink 2003).

Haggard & Eimer 1999, Haggard & Magno 1999, Malenka *et al.* 1982). Moreover, there may be a more general or basic disruption of neuronal processes that affect not just the sense of agency for motor action, but disrupt the sense of agency for cognitive experience (resulting in symptoms of thought insertion). The sense of agency for thought may depend on the anticipatory aspect of working memory, something that may also malfunction in schizophrenic subjects with delusions of control (see Daprati *et al.* 1997, Franck *et al.* 2001, Singh *et al.* 1992, Vogeley *et al.* 1999).¹⁰

Although it is important to sort out the specific nature of the neurological problem, for purposes of this essay one can say that whatever is the precise nature of the neurological disruptions responsible for delusions of control or thought insertion, some such processes clearly generate a first-order phenomenal experience that lacks a sense of agency. The neurological problems occur in very basic, primary mechanisms for motor control, or working memory, and generate a first-order *sense that I am not the agent* of my movement or thought. And this is just the phenomenology reported by the schizophrenic patient. Clearly, both the supposition of neuroscience and the most reasonable explanations of how experience is generated suggest that what phenomenal experience is like depends (at least in part) on the proper functioning of the brain. In the case of the schizophrenic who suffers from delusions of control and thought insertion, neurological problems generate a first-order experience that does not include a sense of agency for certain actions and thoughts. The experience, summarized in a first-person, albeit abstract way, is this: “I am not the agent of my movement or thought.”

If this in fact is an accurate characterization of schizophrenic experience at the level of first-order phenomenal consciousness, then in some respects it seems clear that when a second-order report on such experience is either confused or quite clear about the lack of agency for a particular movement or thought, it is not mistaken. The subject correctly reports her experience, that she experiences herself as the subject (owner) of,

¹⁰ Even if we were reticent to accept the schizophrenic’s own reports of such experience, of which there are many, behavioral studies support this interpretation. A simple but elegant example is provided by an experiment conducted by Frith and Done (1988) which demonstrates that schizophrenic patients are unable to distinguish between self-generated or externally generated sounds. A randomly generated and relatively loud tone will elicit a relatively large response in EEG, but if a normal subject generates the tone himself in a self-generated and spontaneous action of pressing a key, the evoked response will be of a smaller magnitude. One-hundred percent of control subjects showed larger responses to externally-generated than to self-generated tones. In contrast, eighty percent of schizophrenic patients were equally startled (as measured by EEG response) in the two conditions. A significant number of patients surprised themselves, and did not experience self-agency in the production of the tone. More recent experiments on the ability of schizophrenic patients to anticipate events show results consistent with this study (see, e.g. Posada *et al.* 2001).

but not the agent of the movement; she attributes ownership but not agency for the thought. At higher-order levels of introspective report, or reflective metarepresentation, the subject simply reports what she experiences at the first-order level. This is not a mistake, as suggested by Stephens and Graham, but a report of what the subject actually experiences.

This, however, is not the entire story. If the only problem were a lack of agency, then there would be no difference between the phenomenology of passive, involuntary movement, or unbidden thoughts, and schizophrenic delusions. The explicit delusional component of schizophrenic experience involves the subject's misattribution of agency to another person (or machine, or thing). In this regard, a lack of a sense of self-agency does not add up to an attribution of action to others. Here there appears to be two possible explanations.

(a) The cause of the misattribution is based on inferences made at the higher-order level of attributive consciousness.

In this case, the subject introspectively misinterprets her experience as caused by someone else. The lack of a sense of agency is filled-in by a productive narrative. Since the subject has no sense of agency for the action, she makes out that it must be someone or something else that is causing her to act that way. The misattribution may be a supplement or a way to deal with the problems found in experience -- a higher-order, personal-level way of coping with what seems to be the facts. It may be generated in an overactive introspection, motivated by the first-order experience of a lack of agency. This is quite consistent with Graham and Stephens' proposal. They suggest that, in regard to inserted thought, the content of the thought may seem to the subject relevant to context, or appropriately intentional, and thus reflective of agency. But since they are not the agent, someone else must be responsible. The alien nature of the thought is thus a conclusion drawn through higher-order considerations. In a similar way, movements of my body that do not seem to belong to me lead me to a conclusion. "I *conclude* that I am possessed, that my movements are directed by the intentional states of another and express his or her beliefs and desires" (Graham & Stephens 1994: 106-107, emphasis added).

Evidence against this top-down account comes from an examination of the effects of abnormal metarepresentation in pathologies other than schizophrenia. Top-down explanations bestow on the kind of metarepresentational introspection found in schizophrenia the power to generate self-alienation. But if it can be shown that the specific type of higher-order introspective cognition found in schizophrenic patients can also be found in other pathologies that do not involve introspective alienation, then just this kind of introspective cognition would not be sufficient to explain the schizophrenic effects (e.g. the misattribution of agency) manifested in thought insertion and delusions of control. It turns out that one can find pathologies other than schizophrenia (cases of utilization behavior, Anarchic Hand Syndrome (as distinguished from Alien Hand Syndrome (Frith & Gallagher 2002; see reports in Della Sala 2000, Marchetti & Della Sala 1998, Tow & Chua 1998), and obsessive-compulsive disorder) that manifest similar abnormal forms of metarepresentation, but do not involve introspective alienation (see Gallagher 2004a for discussion)).

A second possible explanation can be stated as follows:

(b) Some neurological component responsible for the differentiation between self and other is disrupted, and as a result, some sense of alterity is already implicit in the first-order experience.

In this case, the attribution of agency to another is not the result of a theoretical stance that would force the subject to infer that since he is not the agent, someone else must be, or a supplemental account generated in introspection, the odd result of a productive narrative; rather, it is a genuine report of what is truly experienced. This is not meant to rule out the fact that odd, paradoxical, and wildly delusional narratives are often generated as the illness develops. The initial motivation for such narratives, however, may be shaped by processes that start out as completely rational at the second-order level – that is, a completely correct report of what the subject experiences.

One can find evidence in support of this second, bottom-up explanation of misattribution in the neuroscience of overlapping neural representations. The brain areas that are activated when I engage in specific intentional action turn out to be in large part the same brain areas that are activated when I observe someone else engage

in the same activity.¹¹ A number of researchers suggest that just such overlapping or shared representations may play some part in our ability to simulate the thoughts and attitudes of others (Blakemore & Decety 2001, Chaminade *et al.* 2001, Decety 2002; Jeannerod, 2001). This suggests that if something goes wrong with these overlapping neural functions, this “Who” system (Georgieff & Jeannerod 1998), our own movement or our own thoughts may be experienced at the first-order phenomenal level as initiated by someone else.¹² There is good evidence that this is what happens in some schizophrenic patients. (Jeannerod *et al.* 2003). In such cases, then, not just the lack of a sense of agency, but also the immediate sense of alterity, may be implicit in first-order experience.

Conclusion

Let me conclude with some important qualifications in regard to the bottom-up explanation I’ve been developing. By defending the idea that neurological disruptions may generate problems with the sense of self-agency at the first-order phenomenal level, and that second-order ascriptions may be correct reports of what the subject actually experiences, I do not want to suggest that I have sketched the entire etiological line that leads to schizophrenic symptoms. Obviously there is more complexity both to the brain and to experience than we can take in here. But more than that, I do not want to rule out the possibility that personal-level phenomena may start the ball rolling; that some emotional or intersubjective event may spark the genetically predisposed brain to shift towards a more schizophrenic dynamics. In many cases, some personal-level aspect seems implicated since the schizophrenic often reports that it is some specific person (or machine or thing) responsible for the movement or thought, and that the inserted thought consists of a specific message or content.¹³ In this regard, aspects of

¹¹ Brain mapping experiments (using PET or fMRI) during action, the imaginary enactment of one’s own action, and the observation of another person’s action show activation of a partially overlapping cortical and subcortical network that includes structures directly concerned with motor execution (motor cortex, dorsal and ventral premotor cortex, lateral cerebellum, basal ganglia) and areas concerned with action planning (dorsolateral prefrontal cortex and posterior parietal cortex). In the premotor cortex and the supplemental motor area (SMA), the overlap between imagined and observed actions is almost complete (Decety *et al.* 1997, Grezes & Decety 2001, Jeannerod 1999, Jeannerod & Frak 1999, Jeannerod 2001).

¹² “This relative similarity of neurophysiological mechanisms accounts for both the fact that actions can normally be attributed to their veridical author, and that action attribution remains a fragile process. Indeed, there are in everyday life ambiguous situations where the cues for the sense of agency become degraded and which obviously require a subtle mechanism for signaling the origin of an action” (Jeannerod *et al.* 2003).

¹³ Beyond the issue of misattribution, there are in fact two unresolved problems concerning thought insertion and delusions of control noted in the clinical literature but rarely addressed in the theoretical literature. One is the *problem of the episodic nature of positive symptoms*: for example, the fact that not all but only

Graham and Stephens's account may play a role in explaining why the subject misattributes agency to a *specific* other -- notably, their suggestion about the effect of emotion may be important. And once underway, hyper-reflective introspection may enhance that shift in dramatic ways, producing the frequently hyperbolic narratives and disrupted thought processes of schizophrenia (see Gallagher 2003). In the end, then, a full explanation is likely to involve a combination of these sub-personal and personal factors.

References

- Campbell, J. (1999). 'Schizophrenia, the space of reasons and thinking as a motor process'. *The Monist*, 82 (4): 609-625.
- Chaminade, T., & Decety, J. (2002). 'Leader or follower? Involvement of the inferior parietal lobule in agency'. *Neuroreport* 13 (1528): 1975-78.
- Danckert, J., Ferber, S., Doherty T., Steinmetz, H. Nicolle, D., & Goodale, M. A. (2002). Selective, Non-lateralized Impairment of Motor Imagery Following Right Parietal Damage. *Neurocase* 8: 194-204.
- Decety J, Chaminade T, Grèzes, J., & Meltzoff, A.N. (2002). 'A PET Exploration of the Neural Mechanisms Involved in Reciprocal Imitation'. *Neuroimage* 15: 265-272.
- Decety, J., Grèzes, J., Costes, N., Perani, D., Jeannerod, M., Procyk, E., Grassi, F., & Fazio, F. (1997). 'Brain activity during observation of actions: Influence of action content and subject's strategy'. *Brain* 120: 1763-77.
- Della Sala, S. (2000). 'Anarchic hand: the syndrome of disowned actions'. *Creating Sparks, The BA Festival of Science*. www.creatingsparks.co.uk.
- Daprati, E., Franck, N., Georgieff, N., Proust, J., Pacherie, E., Dalery, J., & Jeannerod, M. (1997). 'Looking for the agent: An investigation into consciousness of action and self-consciousness in schizophrenic patients'. *Cognition*, 65: 71-96.
- Farrer, C., & Frith, C.D. (2001). 'Experiencing oneself vs. another person as being the cause of an action: the neural correlates of the experience of agency'. *NeuroImage* 15: 596-603.
- Fish, F. J. (1985). *Clinical Psychopathology: Signs and Symptoms in Psychiatry*, ed. M. Hamilton. Wright.
- Fournier, P. and Jeannerod, M. (1998). 'Limited conscious monitoring of motor performance in normal subjects'. *Neuropsychologia* 36: 1133-1140.
- Franck, N., Farrer, C., Georgieff, N., Marie-Cardine, M., Daléry, J., d'Amato, T., & Jeannerod, M. (2001). 'Defective recognition of one's own actions in patients with schizophrenia'. *American Journal of Psychiatry*, 158: 454-59.
- Frankfurt, H. (1976). 'Identification and externality'; in A. O. Rorty (ed). *The Identities of Persons* (pp. 239-51). Berkeley: University of California Press.

some of the schizophrenic's thoughts are experienced as inserted thoughts. That this is the case is clear, not only from empirical reports by patients, but by logical necessity. The subject's complaint that various thoughts are inserted depends on a necessary contrast between thoughts that seem inserted and those that do not seem inserted -- and at a minimum, the thoughts that constitute the subject's introspective complaint cannot seem inserted. If all thoughts were experienced as inserted by others, the subject would not be able to complain "in his own voice" so to speak. The second problem is *the specificity of positive symptoms*. In this regard, in cases of thought insertion, specific kinds of thought contents, but not all kinds appear to be thought inserted. For example, delusional experiences are sometimes associated with specific others. A schizophrenic may report that thoughts are being inserted by a particular person, or that they are always about a specified topic. In auditory hallucination the voice always seems to say the same sort of thing. I argue elsewhere that explanations that remain totally on the sub-personal level will not be able to address these two problems (Gallagher 2004b).

- Frith, C. D. (1992). *The Cognitive Neuropsychology of Schizophrenia*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Frith, C. D. (2004). 'Comments on Shaun Gallagher'. *Psychopathology*, 37: 20-22.
- Frith, C. D., & Done, D. J. (1988). 'Towards a neuropsychology of schizophrenia'. *British Journal of Psychiatry*, 153: 437- 443.
- Frith, C. & Gallagher, S. (2002). 'Models of the pathological mind'. *Journal of Consciousness Studies*, 9 (4): 57-80.
- Gallagher, S. (2003). 'Self-narrative in schizophrenia'; in A. S. David & T., Kircher (eds.). *The Self in Neuroscience and Psychiatry* (pp. 336-357). Cambridge: Cambridge University Press.
- Gallagher, S. (2000a). 'Philosophical conceptions of the self: implications for cognitive science'. *Trends in Cognitive Science* 4 (1): 14-21
- Gallagher, S. (2000b). 'Self-reference and schizophrenia: A cognitive model of immunity to error through misidentification'; in D. Zahavi (ed). *Exploring the Self: Philosophical and Psychopathological Perspectives on Self-experience* (pp. 203-239). Amsterdam & Philadelphia: John Benjamins.
- Gallagher, S. (2004a). 'Agency, ownership and alien control in schizophrenia'; in P. Bovet, J. Parnas, & D. Zahavi (eds.). *The structure and development of self-consciousness: Interdisciplinary perspectives* (pp. 89-104). Amsterdam: John Benjamins Publishers.
- Gallagher, S. (2004b). 'Neurocognitive models of schizophrenia: A neurophenomenological critique'. *Psychopathology* 37: 8-19.
- Gallagher, S. & A. J. Marcel (1999). 'The Self in Contextualized Action'. *Journal of Consciousness Studies* 6 (4): 4-30.
- Garrens, P. (2001). 'Authorship and ownership of thoughts'. *Philosophy, Psychiatry, & Psychology* 8.2/3: 231-237
- Georgieff, N. and Jeannerod, M. (1998), 'Beyond consciousness of external events: A Who system for consciousness of action and self-consciousness'. *Consciousness and Cognition*, 7, pp. 465–77.
- Graham, G. & Stephens, G. L. (1994). 'Mind and mine'; in G. Graham & G. L. Stephens (eds.). *Philosophical Psychopathology* (pp. 91-109). Cambridge, MA: MIT Press.
- Grèzes, J. & Decety, J. (2001), 'Functional anatomy of execution, mental simulation, observation, and verb generation of actions: A meta-analysis'. *Human Brain Mapping*, 12, pp. 1–19.
- Haggard, P. & Eimer, M. (1999). 'On the relation between brain potentials and the awareness of voluntary movements'. *Experimental Brain Research* 126: 128-33.
- Haggard, P. & Magno, E. (1999). 'Localising awareness of action with transcranial magnetic stimulation'. *Experimental Brain Research* 127: 102-107.
- Hoffman, R. (1986). 'Verbal hallucinations and language production processes in schizophrenia'. *Behavioral and Brain Sciences* 9: 503-517.
- Jeannerod, M. (1999). 'To act or not to act: perspectives on the representation of actions'. *Quarterly Journal of Experimental Psychology*, A 52:1-29
- Jeannerod, M. & Frak, V. (1999). 'Mental imaging of motor activity in humans'. *Current Opinions in Neurobiology*, 9:735-9
- Jeannerod, M. (2001). 'Neural simulation of action: A unifying mechanism for motor cognition'. *Neuroimage*, 14, S103-S109.
- Jeannerod, M., Farrer, C., Franck, N., Fourneret, P., Posada, A., Daprati, E., & Georgieff, N. (2003). 'Action recognition in normal and schizophrenic subjects'; in: T. Kircher & A. David (eds.). *The Self in Schizophrenia: A Neuropsychological Perspective I* (380-406). Cambridge: Cambridge University Press.
- Junginger, J. (1986). 'Distinctiveness, unintendedness, location, and non-self attribution of verbal hallucinations'. *Behavioral and Brain Sciences* 9: 527-28.
- Lambie, J. A. & Marcel, A. J. (2002). 'Consciousness and the varieties of emotion experience: A theoretical framework'. *Psychological Review* 109 (2): 219–259.
- Malenka, R. C., Angel, R. W., Hampton, B., & Berger, P. A. (1982). 'Impaired central error correcting behaviour in schizophrenia'. *Archives of General Psychiatry*, 39: 101-107.
- Marchetti, C. & Della Sala, S. (1998). 'Disentangling the Alien and Anarchic Hand'. *Cognitive Neuropsychiatry* 3(3): 191-207

- Mellor, C. S. (1970). 'First rank symptoms of schizophrenia'. *British J Psychiatr*, 117: 15-23.
- Parnas, J. (2003-in press). 'Anomalous self-experience in early schizophrenia: A clinical perspective'; in T. Kircher & A. David (eds.). *The Self in Schizophrenia: A Neuropsychological Perspective* (217-241). Cambridge: Cambridge University Press.
- Posada, A., Franck, N., Georgieff, N. & Jeannerod, M. (2001). 'Anticipating incoming events: An impaired cognitive process in schizophrenia'. *Cognition*, 81, 209-225.
- Sass, L. (2003-in press). 'Schizophrenia and the self: hyper-reflexivity and diminished self-affection'; in T. Kircher & A. David (eds.). *The Self in Schizophrenia: A Neuropsychological Perspective* (242-271). Cambridge: Cambridge University Press.
- Sass, L. (1992). *Madness and Modernism: Insanity in the Light of Modern Art, Literature, and Thought*. New York: Basic Books.
- Sass, L. (1998). 'Schizophrenia, self-consciousness and the modern mind'. *Journal of Consciousness Studies*, 5: 543-65.
- Sass, L. (1999). 'Analyzing and deconstructing psychopathology'. *Theory and Psychology* 9 (2): 257-268.
- Sass, L. (2000). 'Schizophrenia, self-experience, and the so-called negative symptoms'; in D. Zahavi (ed.). *Exploring the Self* (pp. 149-82). Amsterdam: John Benjamins.
- Sass, L.A. & Parnas, J. (2003) 'Schizophrenia, consciousness, and the self'. *Schizophrenia Bulletin* 29/3: 427-444.
- Singh, J. R., Knight, T., Rosenlicht, N., Kotun, J. M., Beckley, D. J., & Woods, D. L. (1992). 'Abnormal premovement brain potentials in schizophrenia'. *Schizophrenia Research*, 8: 31-41.
- Spence, S. A., Brooks, D. J., Hirsch, S. R., Liddle, P. F., Meehan, J., & Grasby, P. M. (1997). 'A PET study of voluntary movement in schizophrenic patients experiencing passivity phenomena (delusions of alien control)'. *Brain* 120: 1997-2011.
- Stephens, G. L., & Graham, G. (2000). *When Self-Consciousness Breaks: Alien Voices and Inserted Thoughts*. Cambridge, MA: MIT Press.
- Varela, F. J. (1996). 'Neurophenomenology: A methodological remedy for the hard problem'. *Journal of Consciousness Studies*, 3 (4): 330-49.
- Vogele, K., Kurthen, M., Falkai, P., & Maier, W. (1999). 'The human self construct and prefrontal cortex in schizophrenia'. *The Association for the Scientific Study of Consciousness: Electronic Seminar* (<http://www.phil.vt.edu/assc/esem.html>).
- Tow, A. M. & Chua, H. C. (1998). 'The alien hand sign--case report and review of the literature'. *Ann-Acad-Med-Singapore*, 27(4): 582-5

Paper received April 2004; revised February 2007

Rick Grush

Agency, Emulation and Other Minds

An analysis of the explicit representation of one's own agency is undertaken in the context of the emulation theory of representation. On this analysis, the ability to understand oneself explicitly as an agent is the product of a number of mechanisms, including the abilities to implicitly deal with one's own agency and, via imagination mechanisms explained by the emulation theory, to create and maintain a hypothetical surrogate point of view from which oneself appears as an objective agent among others. Coordinating the implicit-subjective and explicit-objective presentations of oneself provides part of the explanation of the paradoxical self-representation as, on the one hand, an agent just like any other – one more element in an objective world; and on the other hand, a special entity unlike any other – the subjective viewpoint that is the unseen viewer of the world. In addition to a first step at an explanation of this phenomenon, novel interpretation of some neurophysiological data is produced that is consistent with this explanation.

1. Introduction

The object of semiotic study – the realm of the meaningful – is exasperatingly complicated, rich and multifaceted. Anyone who has ever tried to make progress in any of these topics knows this all too well. This kind of phenomenon requires for its investigation many kinds of approaches, not only those who choose to focus on some narrow aspect and operate under various simplifying assumptions, but also those who attempt to state in more global terms how the various threads come together to form a

CORRESPONDENCE: Rick Grush, Department of Philosophy, University of California, San Diego. Email: rick@mind.ucsd.edu.

coherent fabric. And at any level of specificity, there is use both for conservative approaches that stick close to already established results and attempt the painstaking task of adding another firmly documented step as well as speculative approaches that chance new ideas, or new ways of putting together old ideas.

In my brief paper, I will address a narrow issue and do so in a relatively bold and speculative way. This issue concerns the biological infrastructure of the most basic forms of our human capacity for inner, mental, or cognitive representation. These forms concern kinds of relatively simple representational capacities that we share with many other animals, and are likely, I believe, to be the foundation upon which the more sophisticated kinds of representational ability – those most distinctive of modern human semiotic potential – are built. Finally, I will speculate a bit as to one of the main enhancements to this sort of capacity, an enhancement involving the representation of agency which initiates the climb from basic representational ability to human semiotics.

In the first section of this paper, I provide a brief overview of an information processing framework that the nervous system implements in order to construct and use representations of its own body and the environment. I use concepts and tools from control theory and optimal estimation to explain the framework, but the qualitative idea is quite clear and requires no mathematical expertise to grasp. It is this: we can conceive of the brain as an entity that controls the body, meaning that it sends control signals to the body and receives feedback from the body and environment as to the result of those control signals. At some point in evolutionary history, the nervous system hit upon a useful strategy. A subsystem in the brain took on the job of trying to mimic the body and environment. That is, it would observe the commands that the brain issued and the resultant feedback obtained from the body and environment, and it would try to learn this mapping. One benefit of such a system is enhanced motor control. The signal speed of neurons is slow enough that during fast movements, feedback from the distant parts of the body can be delayed enough to cause problems. However, a neural circuit, located in the brain right next to the motor control centers, that is capable of providing an estimate of what the body's feedback will be as soon as it receives a copy of the motor command can help to solve this problem. This neural circuit, or emulator, is providing undelayed feedback to the motor centers. The next trick in evolutionary time is for the brain and its motor areas to be able to operate these emulators completely off-line. This allows the nervous system to produce imagery and

to try out sensorimotor hypotheticals before committing to overt action. (By sensorimotor hypothetical, I mean an indication of what sensory consequences *would* result if some motor command *were* executed. It is hypothetical in the sense that at the time that it is evaluated, the motor command is not actually performed, and the sensory consequences are not actually obtained. Of course, the motor command can later be actually executed, and the sensory result would actually be obtained.)

In the second section, I turn to the ways that agency figures in this framework. One way is implicit, as the organism in fact actively engages with the emulator in the same way that it can actively engage the real body and environment. This engagement can be useful for distinguishing between changes in the sensory state that are the result of the organism's own agency and those that are the result of environmental factors. However, in these cases, the emulator is not explicitly representing agency. Nevertheless, the mechanisms of emulation are capable, and perhaps critical, for being able to make agency explicit. This is the second way in which agency figures in the framework.

In the third section, I briefly discuss our capacity to understand others (and our own) minds. This capacity is what we are using when we understand how other agents act in accordance not with how the world actually is, but how they take it to be. Using the account of explicit agency as a backdrop, I briefly describe not only how the emulation theory can underwrite the so-called simulation theory of mind, but also how it can provide a novel interpretation of the mirror neuron phenomenon.

2. The Emulation Theory

The Emulation Theory of Representation is a hypothesis concerning the biological information processing structure of the most fundamental forms of cognitive representation. It has the advantage that it can explain how meaning can arise naturally from evolutionary processes whose initial aim was better motor control. The framework is easiest to introduce through some elementary control theory. My discussion here will be extremely brief, and often quite qualitative. Readers who are after more detail about the theory itself or would like a description of and references to research that support the emulation framework should start with Grush (2004).

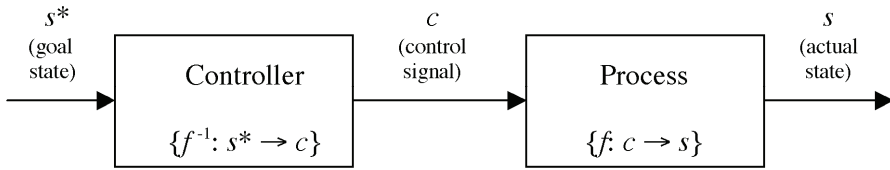


Figure 1. Open-loop control.

The most fundamental frameworks in classical control theory are closed-loop and open-loop control. Open-loop control is illustrated in Figure 1. In an open-loop scheme, a controller controls another system, called the plant or process, by being given a specification of the goal state that the process should go into, and the controller determines a control signal, or sequence of control signals and sends it to the process. The process is then causally influenced by these control signals, and, if they are appropriate, it will exhibit the goal behavior. In a closed-loop control system (Figure 2), the controller is given not only a specification of the goal behavior of the process, but it also gets feedback from the process as the control episode unfolds. It uses this feedback to adjust the control sequence.

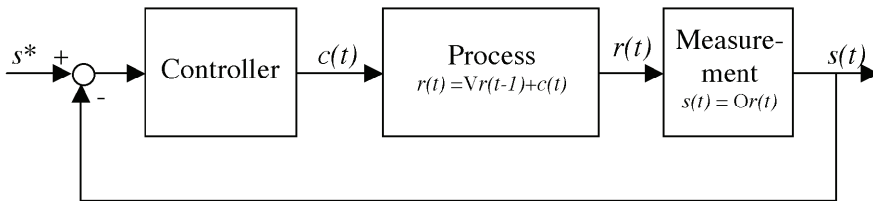


Figure 2. Closed-loop control.

A canonical example of a closed-loop control system is a thermostat. You tell the thermostat what the goal state of the process is by moving a lever. The goal state is, of course, a specific temperature for the room or house. The thermostat also gets feedback from the room or house (the process) in the form of a thermometer reading. Using both of these parameters, the thermostat issues control signals to the heater that manage to get the process to the goal state.

A canonical example of an open-loop controller is a timer toaster. You set the goal state of the toast by sliding a lever (at least on many early models). The toaster

then uses this information alone to determine a control sequence, in this case an amount of time to keep the heating elements on. It gets no feedback from the toast as the episode unfolds.

There are a number of problems that can arise in control contexts that pose problems for closed-loop schemes. These include deadtime, which is a delay that can occur between the issue of a particular command and its resultant manifestation in the feed back signal, and sensor noise, i.e. random inaccuracies in the feedback signal. A number of more sophisticated control architectures are available that can address these problems. The ones relevant for this discussion are all more or less sophisticated variants of the pseudo-closed-loop control scheme outlined in Figure 3. In this control scheme, the controlling system has more than one component. One component of the controlling system is the controller *per se*. The other component is an *emulator*. An emulator is an entity that models the process (also known as a *forward model* or a *process model*). Depending on context, these models might be known beforehand, or they might be learned from observation of the process's behavior. This emulator can be driven by an efferent copy of the same control signal that also drives the process. Since the real process and the emulator are run in parallel in such a case, their outputs will, of course, match.

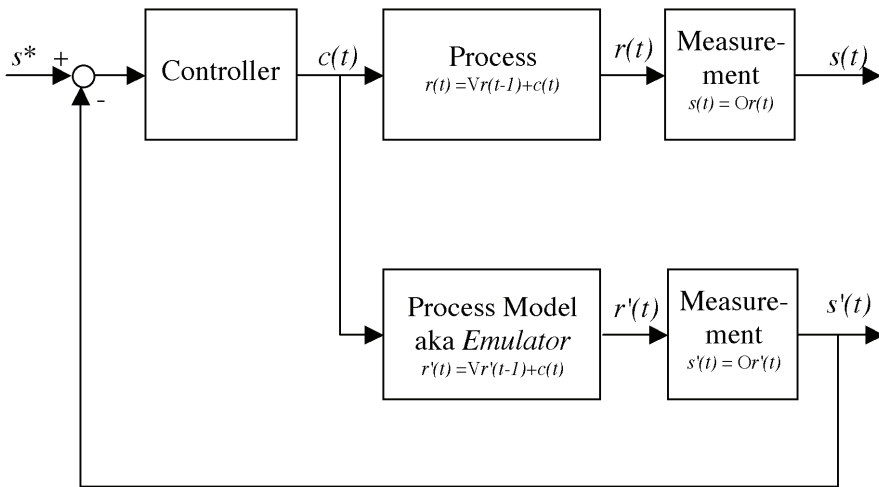


Figure 3. Pseudo-closed-loop control.

This is the basis of the Smith Predictor (not diagrammed here, but see Grush (in preparation), see also Smith 1959), which runs an emulator of the process in parallel to the real process in order to overcome the effects of deadtime (the emulator's output is not delayed). The Smith Predictor is more complicated than my brief gloss, of course. The Kalman Filter (Figure 4, henceforth KF) runs an emulator in parallel with the real process in order to filter noise from the feedback signal. To greatly simplify, the KF combines the prediction of the emulator and the deliverances of the observed feedback signal to produce an optimal estimate of the real, noise-free signal. The controller then uses this filtered signal for its feedback.

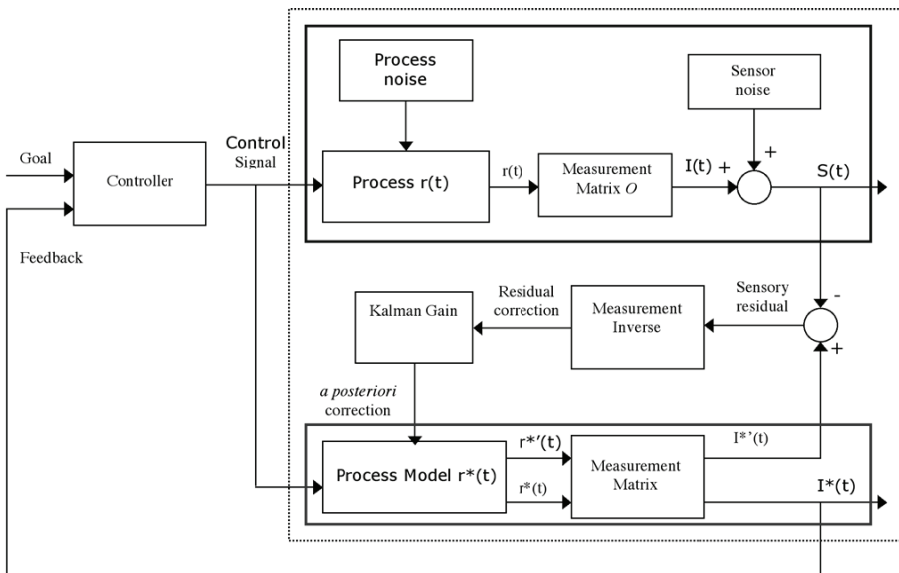


Figure 4. A Kalman filter-based control scheme.

I will very briefly describe here the operation of the KF as shown in Figure 4 by using a familiar example; ship navigation. The ship at sea is engaged in a process that evolves over time, which corresponds to the 'Process' box in the upper third of Figure 3. Its state – location, speed, heading and so forth – at each time is largely determined by three factors: its state at the previous time; any driving force supplied by the ship itself (this is from the 'control signal' box in Figure 3); and disturbances such as ocean currents and winds (the 'Process noise' box). The ship's state is measured, so to speak, at each fix cycle by people taking bearing measurements to known landmarks or stars

(the 'Measurement' box). This is the signal, but it is not perfect. There are limits to the accuracy of these measurements. This inaccuracy can be represented as sensor noise added to what would otherwise be perfect measurements. This corresponds to the production of a noisy signal $S(t)$ by adding noise to a hypothetical perfect signal $I(t)$, in the upper right of Figure 3.

The navigation team's job is to maintain an accurate-as-possible estimate of the ship's state, which can be implemented by making marks on a map. If the team is operating as a KF, it does the following. First, at the beginning of the fix cycle before the sensor information is available (that is, before the people taking the bearing measurements have reported their numbers to the navigation team), the team uses the previous state estimate and knowledge of the driving force to predict where the ship should be. More explicitly, they have a mark on the map that indicates where they think the ship was at the previous fix cycle, and they know what the captain told the person in the engine room and the person at the helm, and so they can produce a new mark on the map that indicates where the ship should be if (i) the previous estimate was accurate, and (ii) the ship behaved as the captain ordered. The result is the *a priori* estimate. This estimate is then subjected to a 'measurement', which is a prediction as to what the sensor measurements will be if the *a priori* estimate is correct. For example, if the ship is at location X , then when the bearing takers report their numbers those numbers should be Y and Z . This expected signal is then compared to the real signal (the numbers that actually get reported by the people taking bearing measurements), and the difference between them is the sensory residual.

The team then uses this residual to update its estimate of the ship's state. The team expected the bearing numbers to be Y and Z , but in fact they were $Y+b$ and $Z+c$. Just as the team expected the bearing numbers to be Y and Z if the ship was at location X , they can determine that bearing measurements of $Y+b$ and $Z+c$ correspond to a ship location of $X+a$. So which is right? The navigation team's expectation that the ship is at X or the bearing measurements that indicate it at $X+a$? Neither is completely trustworthy, since the ship may not behave exactly as the navigation team expects, and the people taking the bearing measurements are not perfectly accurate either. What is done is that some fraction of the sensory residual is applied to the *a priori* estimate in order to produce the *a posteriori* estimate. The exact fraction used depends on the relative size of the sensor noise and the process disturbance. If we assume that the a

priori prediction and the bearing measurements are about equally reliable, then we could simply take the half-way point as the best estimate: $X+(a/2)$.

To apply this to biological contexts, the idea is that the brain constructs and maintains a number of process models or emulators of the body and environment and uses them for a number of purposes. For example, there is evidence (see Miall *et al.* 1993) that the cerebellum maintains an emulator of the body, and particularly the musculoskeletal system, and runs this emulator in parallel with the body during motor control. The purpose is that for fast goal-directed movements, feedback from the body can be slow enough that depending on it in a closed-loop manner could cause problems. The emulator is run in parallel, but its feedback is not delayed, and so its feedback can be used by the motor centers.

There is evidence (see Grush 2004) that motor imagery is subserved by such processes. Motor imagery is the imagined feeling of moving one's body. If there is an emulator of the body that is operated in parallel with the body, a more sophisticated system might be able to run the emulator completely offline in order to produce mock kinesthetic information. This would be analogous to the navigation team simply operating their map entirely off-line, marking out what would happen if various kinds of commands were sent to the engine room and helm and what the corresponding bearing measurements would be without actually moving the real ship.

A similar mechanism can account for visual imagery. On this model, visual imagery is the brain operating an emulator of the motor-visual loop, a system that can anticipate how the visual scene will change as a function of movements. The basic idea is that imagery involves something like operating a flight simulator. The brain's emulator of the body and environment is like a body/environment simulator that the motor centers can operate in much the way that a pilot can operate a flight simulator.

The KF takes this a step further and explains how an emulator that is run in parallel with the real body and environment can provide expectations concerning what will be perceived as specific motor actions are performed. These anticipations are used by the perceptual system to prime them for the likely location of objects, edges, etc. Also, they can be used to fill in missing information. This will be spelled out a bit more in the next section.

3. Emulation and agency

3.1 Implicit agency

The notion of agency is central to the emulation framework as I have described it, for it is a means of allowing the consequences of an organism's own agency (in the simplest case, its own motor commands), together with knowledge of how the body and environment work, to produce anticipations of what will happen. Furthermore, it allows for the evaluation of sensorimotor hypotheticals, which permit the organism to evaluate what the consequences of its agency would be if it were to exercise it in a certain way.

It is crucial to distinguish an *implicit* understanding of, or representation of, something from an explicit understanding of, or representation of, that thing. Perhaps the most familiar example comes from language. All native speakers of a language have an implicit understanding of an enormous range of complex norms that govern the correct use of language (I mean this to be phrased in such a way that any linguist, generative, cognitive, functional, whatever, will agree with it). But few speakers have explicit knowledge of any of these norms. Arguably none do, except perhaps those who have studied linguistics, and even then there are many gaps and likely many errors.

In the same way, an organism whose nervous system is implementing the emulation framework has an implicit understanding of its own agency. This implicit understanding is manifest in the distinction drawn between the control signal that the organism itself issues, the effects of the process, and any process noise. In the ship navigation case, this is the difference between the ship moving forward because the engines are driving it, and the ship moving forward because it is pushed by a strong wind. Such an organism *has* agency, in some sense at least,¹ even if it does not explicitly represent its agency *as* agency. In this section I will discuss the nature of the implicit agency that the emulation framework explains. In Section 3.2, I will turn to the capacity for explicit representation of agency.

The organism's own agency comes into play in the emulator in the form of a copy of the motor command. The emulator itself is capable of maintaining a representation of the situation without a motor command, of course. The emulator

¹ In some areas of research 'agency' is simply defined in such a way that it requires explicit cognizance of oneself *as an agent* in order to have it. In this stronger sense of 'agent', the organisms I am here describing would be self-movers, but would not be *agents*. Only organisms that I try to describe in section 3.2 would count as agents in this stronger sense. What terms we use are unimportant as long as we keep our usage straight.

would then represent what would be happening in the real environment or body if there is no activity by the organism itself. A pilot in a flight simulator does not need to actually operate any levers or move the control stick – in such a situation the simulator will simulate what the aircraft would do without any commands. So a good emulator of use in perception will anticipate how things will evolve even when that evolution is not driven by the organism itself.

Crucially, the emulator needs to be able to handle the case where the organism does act. Part of being able to represent the world as a stable realm even though the sense organs are constantly bombarded by changes is for the organism to be able to cancel out the effects of its own actions. The following example should make this clear.

Every time I move my eyes, the entire scene projected onto my retinas shifts. Yet I do not perceive the world as sliding violently around as I move my eyes. Rather, the world seems entirely stable. How can this be accomplished? A clue comes from a remarkable phenomenon first hypothesized by von Helmholtz (1910) and discussed and verified experimentally by Ernst Mach (1896). Subjects whose eyes are prevented from moving and who are presented with a stimulus that would normally trigger a saccade (such as a flash of light in the periphery of the visual field) report seeing the entire visual scene momentarily shift in the direction opposite of the stimulus. Such cases are very plausibly described as those in which the visual system is producing a prediction – an *a priori* estimate to use the terminology of the emulation framework – of what the next visual scene will be on the basis of the current visual scene and the current motor command. Normally, a motor command to move the eyes to the right will result in the image that is currently projected on the retina (and hence, fed to downstream topographically organized visual areas of the brain) to shift to the left. And some region in the nervous system is apparently processing a copy of this driving force and producing an anticipation of just such a shifted image. This anticipated image is so strong that subjects actually report seeing it briefly. Typically, such a prediction would provide to specific areas of the visual system a head start for processing incoming information by priming them for the likely locations of edges, surfaces, etc. This *a priori* prediction would be largely confirmed, and seamlessly absorbed into ongoing perception. Normally, these images are not noticed, but they are ubiquitous.

Less than one hundred years after Mach published his experimental result, Duhamel, Colby and Goldberg (Duhamel *et al.* 1992) published findings that seem to

point to the neural basis of this effect. They found neurons in the parietal cortex of the monkey that remap their retinal receptive fields in such a way as to anticipate immanent stimulation as a function of saccade efference copies.

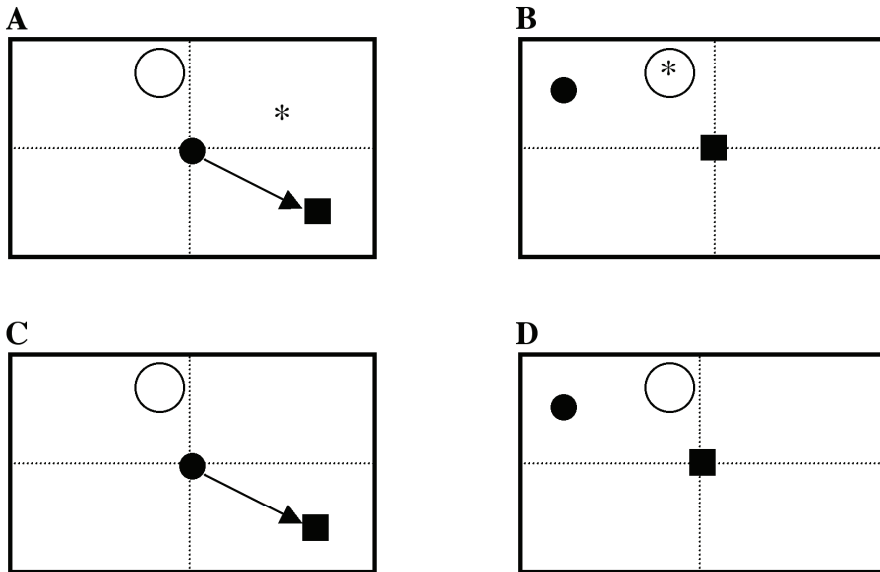


Figure 5. Anticipation of visual scene changes upon eye movement. See text for details.

The situation is illustrated in Figure 5. Box A represents a situation in which the visual scene is centered on a small black disk. The receptive field of a given PC cell (parietal cortex) is shown in the empty circle in the upper left quadrant. The receptive field is always locked to a given region of the visual space, in this case above and just to the left of the center. Since nothing is in this cell's receptive field, it is inactive. The arrow is the direction of a planned saccade, which will move the eye so that the black square will fall in the center of the visual field. Before the eye movement, there is a stimulus, marked by an asterisk, in the upper right hand quadrant. This stimulus is not currently in the receptive field of the PC neuron in question, but it is located such that if the eye is moved as planned so as to foveate the square, the stimulus *will* fall in the cell's receptive field, as illustrated in Box B. The Duhamel *et al.* finding was that given a visual scene such as represented in Box A, if an eye movement that will result in a scene such as that in Box B is executed, the PC neuron will begin firing shortly after the motor command to move the eye is issued, *but before the eye has actually moved*. The PC neuron appears to be

anticipating its future activity as a function of the current retinal projection and the just-issued motor command.

The control condition is shown in Boxes C and D. In this case, the same eye movement to the square will not bring a stimulus into the receptive field of the neuron, and in this case the neuron does not engage in any anticipatory activity. (Or, more accurately, it *does* engage in anticipatory activity, and what it is anticipating is that nothing will be in its receptive field.) This is a neural implementation of the construction of an *a priori* estimate. The PC neuron is predicting that the sensory result of the movement that has just been commanded, but not yet executed, will be a stimulus that will cause it to fire.

A full set of these neurons, covering the entire visual field, would explain the Helmholtz phenomenon. Such a set of neurons would anticipate the next visual scene that would result from an eye movement. And they also provide the clue as to how the capacity of a visual emulator to anticipate the results of the organism's own agency helps it to stabilize the visual scene. Items that project onto the retina *as predicted on the basis of the organism's own motor action on its sense organs* have not really moved. Their apparent motion was just the result of the organism's own action. Any item that has changed location in a way that was not predicted has really, objectively, moved.

Much more could be said about this, but these few remarks will have to suffice. The point for now is that the capacity of an emulator to be able to selectively use information about the organism's own actions is critical for its ability to stabilize an objective world as being the content of its perceptions. While this is not an explicit representation of its own agency as such, it should not be underestimated. Without it, the organism would not even be able to represent a stable world, let alone be able to make use of more sophisticated kinds of explicit representation of its own agency.

3.2 Explicit agency

In the emulation framework, some entity is represented explicitly if it is one of the elements represented within the emulator. An emulator, as a model of the process, is itself a collection of elements that correspond to elements in the represented domain. An emulator of the musculoskeletal system of the sort that can be used to represent the body's own movement will have elements that correspond to physical states of the body, such as the angular inertia of the arm, the tension on the quadriceps, and so

forth. An emulator of the environment will have entities corresponding to objects and surfaces, and it will represent the location and other properties of these objects and surfaces. The emulator used by the ship navigation team is the map and the marks on it, which explicitly represent the ocean surface, the location and orientation of the ship, land masses, and so forth. These entities are explicitly represented.

A simple version of an emulator used by the brain to represent objects in the environment will represent the location of these objects in egocentric spatial terms. The tree is represented as being straight ahead, for example. And the processing of an efferent copy allows the organism to be able to anticipate that the tree will be a bit closer when the action to *move forward* has been executed. But *this* doesn't require the emulator to represent the organism itself, or represent its agency as agency. Note that in my diagram of the Duhamel *et al.* result in Figure 5, the organism itself is nowhere represented. Nor is the emulator. The only things explicitly represented are the black disk, the black square, the asterisk stimulus. The eye does not normally see itself, and in the same way, perceptual emulators do not normally represent themselves.

If an organism has the capacity to take its perceptual emulator completely off line, then it can not only engage in purely fanciful imagery, but it can also construct representations of what the environment will look like from other points of view. I can, for example, imagine what it would look like to get up out of this chair, walk over to the door, turn around and look back at the table at which I am currently working. I can also imagine seeing myself sitting at this desk. In so doing, I can take up the point of view of another representer, one that is explicitly representing me. Let us call this sort of environment from an imagined point of view an alter-egocentric representation. It is still egocentric in the sense that what I am imagining is how the environment would look like from some other point of view, the point of view of an alternate ego, and from that other point of view things would be egocentrically located. But since that other point of view is not where I, the *real* ego, really am, it is not really *egocentric*.

So while during normal perception I do not explicitly fall within what I am representing (except insofar as I can see my own hands, and the edge of my nose, etc.), I can produce, via emulator-provided imagery, a mock experience from some alternate point of view that does include an explicit representation of me.

Now, with yet a bit more sophistication, it is possible for me to run these two emulators in parallel as I actually move about. The first emulator is the one already

discussed, the one in which my own motor commands are used to anticipate how my egocentrically specified environment will change as I move. This is part of what lets me perceive the world as stable. The tree gets closer and takes up more of my visual field as I issue the ‘walk forward’ motor command, etc. But the very same motor commands can be processed by the alter-egocentric emulator in a novel way. It can be used to predict that when I actually issue the motor command ‘move forward’, then my body *as viewed from this alter-egocentric point of view* will move closer to the tree. Note the difference here: the egocentric emulator anticipates that *the tree gets closer*, that it will take up more of my visual field, and so forth; the alter-egocentric emulator anticipates that *the explicitly represented body will get closer to the tree*.

From this point of view, I am not only explicitly representing my own body, but I can use information about the actions I will perform to anticipate how my body will move in the environment, at least as perceived from this alternate point of view.

This combination of two emulators working in tandem now supplies the wherewithal to represent myself explicitly as an agent. While the emulator that is tied to my own sensorimotor situation represents a stable egocentric environment, it does not include myself explicitly as an ego in that environment, and hence, in some sense does not represent either the environment or myself objectively. However, the alternate point of view does exactly that. From within that representational scheme, I am represented explicitly, as an object among others, perhaps as a creature like others of my own species. The motor command that is copied to the egocentric emulator and simply updates it cannot be treated the same way in the alter-egocentric emulator. My actual ‘move forward’ motor command should not operate on this emulated scenario in the same way it operates on my egocentric emulator. It should not move the imagined point of view forward, for example. Rather, the motor command must be disengaged from this alter-egocentric emulator, allowing this imagined point of view to remain stationary even when my real point of view is moving.

Although the motor command cannot be used in the same way by the alter-egocentric emulator, it cannot be completely ignored, for it does make a difference. It is what allows this emulator to anticipate that my body will move closer to the tree. The motor command must be represented as an element *in the emulated realm* – as *the motor command, the intention, of that organism there, the one that is moving closer to the tree* (the one that is actually me).

Now, of course, this capacity to represent an organism as the kind of thing that moves itself around an environment is not necessarily limited in its application to me. Once this capacity is available, it will at the same time allow for the representation of other organisms as agents like myself. That is, as beings that move themselves around with intentional acts and that represent their environment in an egocentric sort of way. Their movements won't be as predictable as my own, but they will be classifiable as the same type, and their unpredictable self-movement becomes intelligible as agentic action on analogy with my own actions.

4. Theory of mind and mirror neurons

Our capacity to understand ourselves and others as agents – as organisms that are within an objective world, that act on and represent that world – is tremendously sophisticated, and even if the extremely speculative story I have told here is correct, it would require a great deal of elaboration in order to be a viable theoretical position. Some of that elaboration has been done elsewhere. In Grush (2004), I delve into much more detail concerning the emulation framework itself, providing much more complete discussion as well as references to work in neuroscience and cognitive science that supports the framework. In Grush (2000), I discuss in more detail the sorts of mechanisms involved in spatial representation and our capacity to entertain alter-egocentric points of view.

I will now say a few words about one strand that I first discussed in Grush (1995) and which has been taken up more recently by a number of researchers, most explicitly by Susan Hurley (2006). The strand I will outline here though differs in some respects from my own earlier account, and differs more substantially from Hurley's. This is the connection between control theoretic concepts as involved in the emulation theory and what has come to be known as the *simulation theory* of mind. This is an account of what it is that underlies our capacity to see others as having minds, or more specifically as seeing the behavior of others as being a reaction not to how the world is, but how they represent the world as being. The canonical experiment (Wimmer and Perner 1983) in child psychology involves children watching a scenario in which a puppet named Maxi has some chocolate and places it in one location in the house, say Location A. Maxi then leaves, and another agent comes in and moves the chocolate to Location B. Both locations are out of sight. When Maxi returns, children are assessed as

to whether they expect Maxi to look for the chocolate in Location A or Location B. After a certain age children have no problem recognizing that Maxi will look at Location A, since that is where Maxi believes the chocolate to be. Younger children fail, however, and expect Maxi to look where the chocolate actually is.

The phenomenon itself seems beyond question. Three issues that are hotly debated are whether this capacity is innate and only manifests itself at a certain time or is learned; the exact age at which it manifests itself; and what this ability consists in. I am concerned now only with the last of these. There are many positions, but the two major players are the *theory theory*, and the *simulation theory* (for a slightly old, but still very good, introduction, see the contributions in Carruthers and Smith 1996). The theory theory maintains that children have a theory of mentality, a theory that posits a belief-desire psychology and other supporting mechanisms, and the child employs this theory in explaining the behavior of others. The simulation theory maintains that children do not have an explicit theory in this form, but rather simulate the other agent. They put themselves imaginatively in the other agent's situation and assess what they would do in that situation. The child after the critical age is able to put herself in Maxi's shoes and realizes that she would look in Location A.

The simulation theory credits children with exactly the capacity to produce a surrogate point of view, and so the emulation theory of representation is a natural ally of the simulation theorist. What I want to do though is to extend this hypothesis in the following way. The hypothesis concerning what underlies an explicit representation of agency that was outlined in the previous section is quite a bit more involved than the mere 'simulation' of another point of view. It involved the capacity to simultaneously maintain and coordinate two different representational structures: the egocentric point of view and the simulated alter-egocentric point of view. I want now to exploit this additional structure in order to offer an explanation of a phenomenon, tentative of course, that differs from the existing explanations usually given of this phenomenon by proponents of the simulation theory.²

The phenomenon is mirror neurons (Rizzolatti *et al.* 1996). These are single neurons in premotor cortex that become active either when the agent performs some action, or when the agent observes another agent performing the same action (the

² I won't pause to describe these other theories. For a useful and detailed discussion, see Hurley and Chatter (2005).

single cell recordings are done on monkeys, and the neurons have been universally dubbed the ‘monkey see monkey do’ neurons). The existence of these neurons seems to suggest some sort of deep connection between one’s own actions and the actions of others, but what is that connection exactly? Here is a brief speculative proposal (one I will expand upon in Grush (in preparation)).

Recall the details of how the two emulators work in tandem when I am conceiving of my own actions as objective. There is one emulator that is maintaining an egocentric representation of the environment. This emulator includes an implicit representation of the organism’s agency in that the organism’s actions drive the emulator to, for instance, anticipate the consequences of self-actions. The emulator must be capable of processing the organism’s own actions.

The second, alter-egocentric emulator represents the environment from a surrogate point of view. This emulator must, recall, work knowledge of the agent’s own actions into its representational manifold differently from the first. This emulator represents the agent itself as an entity in the environment. And so the agent’s own actions alter the structure of this second emulator by changing how this agent (the self represented from the surrogate point of view) will look as a function of the actions it is about to perform. From the surrogate point of view, the self *is* an other, and is represented as moving closer to the tree – a situation that is represented in the same way as would be the case if was watching some other person walk towards a tree. And it is the understanding of oneself simultaneously from the inside and the outside that underwrites the agent’s explicit understanding of itself as an agent.

This is a possible explanation for the functioning of mirror neurons. In order for an organism to properly coordinate these two representational manifolds, it must have the wherewithal to allow an efference copy of one of its own actions to simultaneously influence its representation of the egocentric environment and also at the same time represent its representation of the scene from the alter-egocentric point of view.

If this is correct, then the mirror neurons are, in a sense, not mirror neurons at all. They have exactly one function: to fire when there is a representation of another agent performing some action. Such a representation is present when the monkey observes another animal perform the action, of course. There the representation is a straight-forward perceptual representation. However, when the monkey is itself performing an action, there is also maintained in its cognitive system a representation

of ‘another’ agent performing an action. In this case, this representation is a mock perceptual representation, the situation as perceived by an alter-ego. And this alter-ego perceives the monkey itself as another agent. The mirror neurons are the crucial link between the agent’s implicit representation of itself, and its capacity to represent itself explicitly and objectively as an agent.

5. Discussion

The proposals I have outlined have suffered from the twin failures of being very schematic and very speculative. This is a powerful combination, since often the only way to save speculative proposals from being vacuous is through detail. But there is another way for a speculative proposal to have value: for it to at least provisionally lay out a potentially powerful synthesis and novel explanation of some range of phenomena. If there is a saving grace available for the present essay, this is it. So it will perhaps be worthwhile for me to say a few words about the nature of the synthesis proposed.

The proposal builds on the emulation framework, which has been independently motivated, and for which a great deal of supporting evidence exists (see Grush 2004, for much more detail). And it also builds upon a worked out account of the capacity for objective spatial representation of the sort embodied in cognitive maps (see Grush 2000, for more detail). It promises to expand this latter account into one that provides the material for not just objective spatial representation, but objective and explicit representation of the agent itself as a locus of agency (this was briefly outlined in the second section). It shows how the emulation framework can provide the information-processing infra-structure for the simulation theory of mind. And finally, it does this last task in a way that provides a novel account of the role played by mirror neurons.

References

- Carruthers, P. & Smith, P. (eds). (1996). *Theories of Theories of Mind*. Cambridge, MA: Cambridge University Press.
- Duhamel, J.-R., Colby, C., & Goldberg, M.E. (1992). ‘The updating of the representation of visual space in parietal cortex by intended eye movements’. *Science* 255(5040):90-92.
- Grush, R. (1995). *Emulation and cognition*. UC San Diego Doctoral Dissertation. UMI.
- Grush, R. (2000). ‘Self, world and space’. *Brain and Mind* 1(1):59-92.
- Grush, R. (2004). ‘The emulation theory of representation: motor control, imagery and perception’. *Behavioral and Brain Sciences* 27(3):377-396
- Grush, R. (in preparation). *The Machinery of Mindedness*.

- Hurley, S. (2006). 'Active Perception and Perceiving Action: The Shared Circuits Hypothesis'; in T. Gendler and J. Hawthorne (eds.). *Perceptual Experience*. Oxford: Oxford University Press.
- Hurley, S. & Chatter, N. (eds.) (2005). *Perspectives on Imitation: From Mirror Neurons to Memes*. Cambridge, MA: MIT Press.
- Mach, E. (1896) *Contributions to the analysis of sensations*. Open Court Publishing.
- Miall, R.C., Weir, D.J., Wolpert, D.M., & Stein, J.F. (1993). 'Is the cerebellum a Smith predictor?' *Journal of Motor Behavior* 25(3):203-216.
- Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). 'Premotor cortex and the recognition of motor actions'. *Cognitive Brain Research* 3:131-141.
- Smith, O.J.M (1959). 'A controller to overcome dead time'. *Instrument Society of America Journal*. 6: 28-33.
- von Helmholtz, H. (1910). *Handbuch der Physiologischen Optik*, vol. 3, 3rd edition, edited by A. Gullstrand, J. von Kries & W. Nagel. Voss.
- Wimmer, H. & Perner, J. (1983). 'Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception'. *Cognition* 13:103-128.

Paper received April 2004; revised February 2007

Merlin Donald and Lars Andreassen

Consciousness and Governance: From embodiment to enculturation – an interview¹

Introduction

What some researchers have called the second-generation cognitive revolution consists largely of the study on the embodiment of mind; that is, the bodily basis of meaning. This is an important step toward the inclusion of human meaning in an evolutionary framework. However, placing the human spirit in an evolutionary context also means that we inherit some of the explanatory difficulties of theoretical evolutionary biology. The reluctance of the humanities to try to understand human meaning in an evolutionary framework is based partly in the darker aspects of phrenology and eugenic thought to which biological theories gave rise in the last century. This reluctance might also be due to the difficulties of explaining the uniqueness of human culture and human achievements, such as language, writing, and art, which took place only recently (40.000 years ago). One main obstacle in explaining the evolution of human culture seems to be that it is not enough just to list a few genetic changes.

Professor Merlin Donald has suggested that in order to explain the unique achievements of human cultures, it is not enough to consider consciousness or the human mind as mere by-products of our brain's evolution. We also, or even mainly, have to consider the mechanisms of what Donald calls “enculturation”, which is the ability to create external symbols, to hand down habits that ground the unique cultural

¹ The interview took place May 8th 2003 at the Center for Semiotics, University of Aarhus, Denmark, following two guest lectures by Professor Merlin Donald.

development of human societies. Thus, in the last 30.000-40.000 years, human cognition and the modern mind have been shaped by the dynamic interaction between brains and culture.

One great advantage to Donald's theory of enculturation is that it accounts for how the executive functions of consciousness have come into existence in evolution. Donald argues that the executive functions of consciousness depend heavily on cultural artifacts and only partly on neural tissue. Human beings are able to use symbolic means to externalize and materialize their thoughts. This not only makes it possible to pass on the progress of one generation's labor to the next, but also provides our minds with new input. Thus, symbols function as expansions of the mind, and as a corollary, the structure within different classes of external symbols also provides structure to the mind. By externalizing some of its operations and joining itself to a network, consciousness becomes the generator of cultural and cognitive development, and that development in turn deeply affects consciousness itself.

Donald's theory thus provides a new way to consider different aspects of volition, decision-making, and free will in human beings; phenomena that have always been intuitively evident to us, but still seem very hard to explain in a scientifically and theoretically consistent way.

Consciousness and agency

Lars Andreassen: The philosophers and scientists that you call "hardliners" claim that consciousness does not seem to possess much generative power: This is especially evident when it comes to perceptual-motor activities such as driving, musical performance, speaking, or sports. In such situations, it actually seems crucial not to be explicitly aware of your actions. These scientists do not necessarily consider consciousness an illusion; rather they think that it functions in a secondary role: We might be conscious of what we are doing, but this awareness seems to come some time after the initiation of the specific action. A lot of complex behavior seems to be automatic, and consciousness seems to be very limited in its powers. You react strongly against this view on consciousness, even though you do not want to abandon the experimental data that suggests that it is so.

Merlin Donald: I think that agency is definitely initiated in awareness. I do not say that every act is consciously regulated, but I think that especially in anticipatory planning of action and in meta-cognitive monitoring of self-action and of action in others, conscious or deliberate meta-cognitive control is essential; it is an integral part of the mechanism of voluntary action. Ten years ago, Dennett argued that we could take the electro-physiological evidence on motor control and show that the electrical preparatory activity of neurons actually preceded action; and that it preceded our awareness of our intentions to act. The phenomenon he referred to was called the *Bereitschaftspotential* (Readiness Potential) and was discovered by Hans Helmut Kornhuber and Lüder Deecke in the 1960s. They showed that prior to the production of an action, the motor and pre-motor cortex are polarized, and we see a very strong electrical signal coming out of that part of the brain. This signal emerges as much as a second or a second and a half before the action is actually carried out. Dennett based his arguments on some famous experiments carried out by Benjamin Libet about thirty years ago using the *Bereitschaftspotential*. What Libet did was quite clever. He had his subjects monitor a slow moving clock and register in memory the moment they decided to move a finger. The action they performed was simply lifting a finger. Libet was able to trace the activity of the finger exactly in time, monitoring the muscle that contracted and raised the finger, and the electrical activity of the brain that preceded the muscle contraction. At the same time, he recorded the subjective report of the position of the hands of a clock when the person decided to move. What he found was that the actual electrical preparation for the movement had begun before the person had consciously decided to move. Dennett interpreted this to mean that awareness followed rather than preceded the brain's initiation of action. However, I think that Dennett and others have overinterpreted those results. The electrical activity Libet observed is not necessarily tied to the movement itself, but rather to the anticipation of the decision to move. It is entirely reasonable to expect that prior to a decision to move, there should be brain activity related to it. This is hardly surprising.

Andreassen: What you are saying is that there is activity somewhere in the brain which is related to the conscious decision to move before the specific command from the motor cortex?

Donald: Yes, I do not think that there is any good evidence of consciousness being an epiphenomenal result of prior “unconscious” neural activity. In fact, any conscious decision is itself an aspect of neural activity, and must have a physical instantiation (if we are to be consistent Monists). I tend to see both the conscious decision and its neural correlate as two aspects of the same reality, and I do not buy into the idea that the *Bereitschaftspotential* is evidence for the elimination of consciousness from motor agency. Similarly, I have questioned the validity of all the so-called empirical evidence that Dennett claims as proof that awareness is irrelevant. We have no such proof.

Andreassen: You suggest that different parts or sub-elements of behavior are normally integrated by conscious effort. After some extensive rehearsal, they become automated action-programs and eventually make up a coherent and seamless pattern of complex behavior, such as driving a car or dancing. You call these automated action-programs ‘demons’ – after the AI tradition I suppose – and in one of your lectures today you called them ‘agent systems of the brain.’ If we consider these demons in the light of Libet’s experiment, then the movement of the finger seems to come before the decision is actually made. But only because, at a higher level, the subject already knew that he was going to move the finger, and therefore the demon was already up and running on the basis of a prior conscious decision to activate that particular sequence. Would it then be appropriate to say that you regard automaticity as a by-product or slave system of conscious awareness or integration?

Donald: Yes, it usually is. Conscious rehearsal would not necessarily be the only way automated skill can be acquired, but I certainly think that it is one important way in which consciousness plays an active role in behavior. When we think of agency, we usually think of direct action, but if we extend the definition of agency to include deliberate thought and imagination, then it is evident that conscious regulation plays a role there as well. Luria’s famous case of Zasetzky, which I re-examined in *A Mind So Rare* (Donald 2001:71), is a good illustration of the important supervisory role the executive brain plays in the conscious self-regulation of behavior – even in cases of such tremendously crippling cognitive disorders as Zasetzky’s memory disorder. In fact, there are very good reasons to think that conscious or deliberate planning and self-regulation are the most important factors in behavior. Even when behavior involves

chains of automatisms, these are the product of conscious rehearsal and practice, and always subject to further conscious modification and regulation.

Andreassen: We have some automatic behavior patterns that may be genetically determined or innate. Let us take the example of the fear reaction. While walking in a forest, it is possible to mistake a stick for a snake², or in the city, a lost glove for a rat (which I once did). Fear is processed by the danger-detection system based in the thalamus and the amygdala, which produces an automatic fear-reaction to the stimulus. After a small delay, the visual system then gets the same information and is able to conclude that the stimulus was not a snake or a rat, and the body may then regain its homeostatic baseline. Such a reaction system could be viewed as a demon that is different from those installed by studying to become a concert pianist, or learning to play tennis, or driving a car. Is it possible to make a distinction between these two kinds of demons?

Donald: I would not say that snake phobias are innate in the sense that they are implanted in the brain tissue *per se*. However, humans have a predisposition toward certain types of fear reactions. In other words, some phobias are easily learned, whereas others are not. The evidence seems to be that phobias are learned, but that we are predisposed to learn particular phobias. Thus, it is easier to become afraid of spiders and snakes, than of beautiful women, even though it is also entirely possible to become afraid of the latter, under the right circumstances. There used to be a theory suggesting that there was a deep mammalian fear of reptiles, but I do not know if that is true, because mongooses (*zoo. Herpestes*) eat snakes, which they apparently find delicious. These matters are not simple. Nevertheless, we could have a predisposition to become afraid of creatures that are very different from us, and certainly, spiders and reptiles are very distant from us. We recognize all mammals as being closer to us on an emotional level. It is easier for us to empathize with a dog in pain, than, say, an insect or a fish in pain; it is a matter of emotional proximity. There could be some general predisposition to fear species that are distant, and that has to do with some general bias, rather than an innate fear system. The same applies to many other innate reactions. Rats, for example, have a very sensitive olfactory system, and they can learn in a single trial to avoid

² The example is taken from Joseph LeDoux (1996).

poisons. It has been shown that they can learn this even when the stimulus and the sickness associated with it are separated by several hours, despite the fact that it is almost impossible to teach the same animal other associations involving long delays. There is apparently a natural connection between the gastro-intestinal system and the taste-system, which equips the animal to learn that particular kind of association easily. However, this does not necessarily mean that a complex innate mechanism is involved. It could be a simple bias that allows an animal to be predisposed to learn important things.

Andreassen: The emotional systems, such as the fear system, are very easily conditioned. Would you regard enculturation as a product of emotional conditioning?

Donald: Yes, absolutely. At least, it starts there. I think that human beings are predisposed toward certain types of early emotional relationships. We see it early in ontogenesis in the mother-child bond. During adolescence, under hormonal influence, mating and the opposite sex suddenly become tremendously important. (Many years later, we come out of the trance and wonder what happened.) This is a common pattern in animals where many behaviors are under hormonal control. Humans are not an exception. The same applies to emotional disorders. There are depressive, obsessive, and other negative emotional experiences that may result from a biochemical imbalance in the human body rather than from an obvious outside cause. A person with depressive tendencies may have a chemical bias, but not necessarily a complex built-in mechanism that governs depressive behavior. These biases predispose the person to conditioned emotions. Emotions tend to affect associations in memory, and almost all associations have emotional valence. When we experience a very strong negative emotion, all sorts of negative associations flood into our awareness. Unless we are very careful, we can become obsessive about them and keep going repeatedly over the same type of material. I have no doubt that an underlying cause could be simply an endogenous tendency to tag associations with an emotional value. We do not necessarily have to build in complicated brain modules, or innate reactions, to explain these predispositions.

Andreassen: When the automatized demons are installed in the brain, do they completely take over control and run the actions on their own?

Donald: To a degree, but they are normally monitored in awareness. Error correction is one of the vital functions of conscious processing. Any learned algorithms and learned behavior patterns involve functional rewiring of the brain, in the sense that functional circuits are created that did not exist before. When we learn to speak or to drive, we set up a new functional neural architecture. This is only possible through conscious integration. There is very little evidence of complex learning outside of consciousness. But the conscious mind has no awareness of the brain processes upon which its efficacy depends. This does not imply that those brain processes can be labeled as “unconscious.” Mentality and physicality are two aspects of the same underlying reality.

Learning and representation

Andreassen: In your theory of cultural evolution, you strongly emphasize the importance of awareness, especially in the learning and rehearsal of skills. Nevertheless, awareness is often considered a more or less passive element in our lives. John Lennon for instance sang: “I’m just sitting here watching the wheels go round and round.” Lennon was not a cognitive scientist, but I, and possibly others, feel that learning seems to require effort. I have to focus my attention on the subject matter and not just be aware of it. There seem to be a distinction here: In *A Mind so Rare*, you mention that it is important to “keep in mind the distinction between controlled processing and explicit self-awareness.” (Donald 2001: 256). Is there a difference between these two concepts?

Donald: There is a clear distinction in the sense that these are two different paradigms, and yet they address a common theme: awareness. Controlled processing implies a goal: awareness of what we are trying to do. Truly passive awareness lacks a goal, or at least, a specific goal. Passive awareness gives control over to outside forces. Active awareness tries to engage those forces and control them. For example, in a complicated sensory-motor task, let us say a video game in which we have to move very fast, our attention has to be fully engaged, otherwise our responses would not be fast enough, and we

would fail to reach our goal. It is possible for us to construct absorbing games and situations that demand a lot of controlled processing. Psychologists who have studied controlled processing find many experimental tradeoff effects that are not present in automatic responses. These are attributed to “limited capacity.” Our conscious capacity is limited, and when we are engaged in controlled processing, we cannot tolerate distraction. We trade off accuracy and speed when we need to perform several things at the same time, because we are using our limited conscious resources to focus on a particular task. If the task is very demanding, such as learning how to drive a car, we cannot tolerate any kind of interruption. Whereas, when we are performing the same task in automatic mode, we can tolerate all kinds of disturbances. For example, in the early stages of learning to drive, we might find it catastrophically disturbing to see somebody walk across the street in front of us. However, once we have learned to drive, we can listen to the radio, or carry out a conversation with someone, and still avoid pedestrians without even noticing it as a significant event. In other words, conscious capacity is something that seems to have a limit, and when it is fully engaged in the controlled processing of a situation, the task itself absorbs all that capacity. But when we are performing the same task in automatic mode, we seem to have extra capacity because we are not engaging consciousness as much and can therefore direct the remaining capacity to other things. Human beings can run many different automatic tasks under conscious supervision at the same time. For example, if we are monitoring our speech in automatic mode, we do not have to think about the act of speaking itself, choosing words, monitoring grammar, verifying intended meaning, and so on. When we are speaking our highly automatized native tongue, we can do many different things at the same time as we speak, provided that we have automatic routines for doing them. This is called multi-tasking, and it is possible only if we have automatized some of the things that we are doing.

Andreassen: But, learning is also largely receiving. You seem to include many passive or reactive elements in your theory; for example, you consider mimetic skill as the platform for cultural evolution.

Donald: I do not consider mimetic skills as passive. Mimesis is a very active type of expression, but some of it is reactive in the sense that it is cooperative, communal, and

collective. Many of the behavioral patterns of groups are a result of this conforming tendency, this mimetic predisposition to reproduce behaviors in groups. However, I do not see that as passive. Rather, I regard it as an active attempt at creating expressive read-out. In groups, expressive motor-patterns represent ideas, memories, intention, and thoughts. Mimesis is actively representational. It is a creative attempt to generate a motor pattern that encapsulates or captures a perception of an event. If we “mimic” a hunter slaying a bear, what we are doing is selectively reenacting the event to convey it to other people or perhaps to remember it ourselves. This is a highly conscious production. Performed in groups, such expressions create a theater of public (mostly nonverbal) communication. Most art and most performing arts in particular involve deliberate mimesis.

Social behavior and communication

Andreasen: Mimetic skill brings us to the domain of social interactions. Do you believe that intentional behavior is something that emerges out of iconic or mimetic gesture? It seems that we sometimes react to an unintentional action as if it were an intentional gesture.

Donald: Presumably, that is where mimesis has its evolutionary origins. When observing behavior even non-intentional behavior, in others, we can conclude things about them. Many animals do this very well. However, that does not make such behavior intentional or symbolic in its genesis; it just means that good observers can attribute meaning to behavior. This is different from intentional communication. We can say that human beings are inherently actors. They are truly theatrical in their actions, when *they know that they are communicating* with people by their expressions, body language, gestures, attitudes, postures, and tone of voice. There is a great tendency among human beings to play a role and to expect confirmation from others. We are a conforming species. This strong conformity of human beings is one of the best pieces of evidence for the importance of mimetic expression in human social life. The social domain is largely mimetic in the sense that it is a theater of action collectively encapsulating the events that define the society rather than an unfolding of events *per se*.

That does not mean that mimetic expression is always deliberate or conscious. Automatization occurs here too. Some of the most common theatrical behaviors and

customs become automatic and invisible, for example, the rituals surrounding the consumption of food. In ontogenesis, we learn how to fit in, and how to adjust to different situations mimetically. When we become members of a culture, we learn the body language and appropriate expressions of that culture in specific contexts. We know that we should not behave in the same way when we go to church as when we attend a party, and this kind of adjustment eventually becomes automatic. If we go into a strange culture, it becomes evident to us that we must learn these new cultural patterns of behavior very deliberately and consciously. Even when these patterns have become automatic and routine, we have to invigilate and monitor our own mimetic behavior in novel social contexts consciously; otherwise, we could easily end up doing something entirely inappropriate and scandalize everyone.

The tricky thing is that the same action can become deliberate and intentional under some circumstances, because it is intended to communicate a state of mind to an audience. This is where we see intentionality, as we normally understand it, very clearly. A person who experiences real grief due to the real death of a real person might be reacting involuntarily, but the same behavioral signs could also be a deliberate display of feigned grief that lacks any real emotion. It is not the superficial aspects of the behavior, but the ways in which it is generated, that define whether it is intentional. Of course, a display of grief can contain both of these elements simultaneously.

There is no evidence to show that apes can act deliberately to produce feigned emotional displays. But their expressions are very similar to human expressions, and it is clear that many human expressions are primate in their origins. For example, laughter is characteristic of humans when defined in terms of its social uses; but there are equivalent vocal behaviors in chimpanzees and other primates. They are not exactly functionally similar to human laughter, but they sound the same and involve the same use of the vocal chords and opening of the mouth, as well as some other elements of its use in social contagion, for instance. Humans use this quintessentially primate pattern of behavior, which is innate and somewhat involuntary, in a very deliberate way to communicate in social situations. Laughter can be used to ostracize or punish people. Laughing at somebody can be a deliberate and crushing kind of exclusion from the group.

In social behavior, humans often do many things simultaneously, because the algorithms are quasi-automatic. This allows for very complex social behavior. For

instance, the heroine in Henry James's novel, *The Bostonians*, uses several layers of social deception at once to achieve her ends. I used this interesting example in my book because the complexity of her social maneuvering is so typical of what we do. We use body language to communicate a certain message deliberately, or we speak to people in ways that are dictated by social conventions in order to manipulate their responses, while we remain fully aware of our deception. As good parallel processors or multitaskers, we have to learn many automatic subroutines so that we can generate and monitor the subtleties of social behavior.

Andreasen: Sometimes consciousness seems to make things more ambiguous. For instance, in the conversation you describe in *A Mind so Rare* (Donald 2001: 48), where eight different people, most of them bilingual, but not all speaking the same two (or more) languages, discuss a movie. In the conversation, the interlocutors keep track of a host of major and minor details. In verbal exchanges, we both listen to what our interlocutors say, and read their body language, gestural signs, facial expressions etc. This involves many things. To some extent, it must be very important to be able to rely on “subconscious” processing of special emotional cues, because if consciousness engages too deeply in the interpretation of such cues, they will become much more ambiguous due to our ability to see things from different perspectives. In addition, the timing of expressions and utterances is very subtle, and conscious deliberation can ruin the timing by separating a glimpse in the eye or a certain movement of the lips from the words, and thereby introducing misinterpretation and confusion. So, the question is would we not stumble in our conversations if we could not rely on some sorts of automatic emotional systems to read expressive signs?

Donald: Yes, we read emotions as quickly and as automatically as we parse sentences. It is an interesting question to try to figure out at what level in a verbal exchange consciousness kicks in. I think it operates at many levels simultaneously. It plays a crucial role at a very high regulatory level, almost at the level of supervision. Yet it can intervene at a low level if it wishes. Let us compare the art of conversation with that of a pianist playing the piano. A professional pianist is not thinking about which finger(s) to move next or what part of the melody comes next; otherwise, the interpretational elements of his playing will be disturbed. While he is playing, he must think about the

kind of interpretation he wants to get across to the audience. Yet, if one of his fingers is malfunctioning, his awareness can intervene on the level of specific motor commands. Conscious awareness can move through cognitive space much as a zoom lens moves through visual space.

One thing that is central to the notion of conscious regulation is that it is multilevel; there are hierarchies of attentional control so that attention itself can be automatized at the subroutine level. The best example of this is reading. In reading, the eye moves as it scans the page, the reader detects the ends of the line and goes to the beginning of the next line. The brain performs a series of learned algorithms that regulate the fixations by which the reader takes in print. All these movements result in an input to the brain that would actually look quite confusing if we were to present it on a screen as a series of stimuli. Somehow, in learning to read, the various eye movements have become automatic; but they are still regulated by conscious attention, because the sequential fixations must be selectively processed at a high level for purposes of error detection. Thus, there are semi-automatic attentional hierarchies that place conscious regulation very high up in the system from where it can interrupt, decide, and voluntarily focus on something that may be going wrong on some lower level. For example, when a musician is playing the piano and something sounds wrong in his right hand or even in a certain finger in his right hand, he can suddenly focus on that and move down in the hierarchy to fix the problem. Similarly, in reading, the mind can move up and down the attentional hierarchy, perhaps to rescan a part of the text that was not properly processed in the first place.

This multilevel nature also implies that consciousness is not a unitary phenomenon, or a matter of all or none. Consciousness is a graded hierarchy of up-and-running routines. There is a “vivid core” of sensory awareness and various other active elements and routines, running concurrently under metacognitive supervision. At any given moment, the metacognitive observer can suddenly broaden the reach of awareness or focus it on some particular detail. This theoretical view of conscious regulation is very different from that of the hardliners. Consciousness is not narrow; it has a wider reach, a wider temporal and spatial span, and a more complex structure than what is generally allowed in the laboratory paradigms of experimental psychology.

The externalization of representations

Andreassen: The invention of a variety of refined ways to represent, where mimetics is only one of them, plays a tremendous role in your theory of mind and cultural evolution, especially when you talk about external symbols. You say that symbols do not seem to carry much transformational power in themselves (Donald 2001: 204), and that they seem to have served initially for the reproduction and consolidation of old customs (Ibid.: 306). This seems to suggest that you view our symbolic technologies as organizing devices, more than direct vehicles for creating new meanings. Do you regard new ways of sorting and storing knowledge as a generative activity?

Donald: The idea of external symbol, i.e. memory storage, is a very powerful one, because it allows the human species to transcend time and space, and to escape the nervous system. If we create different varieties of external memory storage, such as written or mathematical representations, this enables not only individuals, but also whole cultures to remember and represent reality far better than they would be able to do within the brain's memory systems alone. A megalithic observatory, such as Stonehenge, can record the times of various astronomical events and track eclipse cycles 52 years in length. The observatory is an external memory device and a collective resource, which of course has to be interpreted by a conscious mind. It does not interpret itself. It is a storage device just like the synaptic patterns in the brain. But, at one point, the signal stored in any memory medium has to be interpreted consciously in order to come alive in the forum of active cognition.

The interactions between individual human consciousnesses and external symbols are very interesting and important to study. However, they complicate the role of conscious thought and creation, because they reflect the structure of each mind externally, to the scrutiny of other minds. This can lead to the dilemma of the overloaded observer who has so many symbolic devices available, such as books, scientific papers, and works of art that he is overwhelmed. On the other hand, these symbols are also powerful tools that conscious individuals can use to reflect on their own past. The writings of Marcel Proust, for example, or those of any autobiographer are deliberately reconstructive and creative. Proust was doing an archaeological excavation into his own past in order to reconstruct it. Typically, when we try to reconstruct memory, we gain a more conscious perspective on our lives and behaviors.

But, external symbols make this much easier. External symbolic devices have enhanced our metacognitive understanding of human life. Indeed, one could argue that the primary function of the “text” in civilization is a metacognitive one. It provides the prototype or archetype against which individual lives are measured.

Andreassen: When symbols or representations are externalized, they also function as environmental inputs, and in that way, our brains react to them as displays.

Donald: Yes, and this is an important point. Our natural display of “natural” biological memory is actually quite poor. If we try to recall what somebody said or what an event was like, we find that the brain stores evanescent and inaccurate information. However, we can greatly enhance the power of memory with external devices, and that is why external symbols are so powerful. They are easier to display, much more stable and permanent, and infinitely reformattable.

With the help of simple external memory devices, early humans were able to calculate planting and hunting seasons, make predictions of weather changes and other important things. Some of the earliest external symbols were apparently portable calendars. For example, there are carved bones about 15,000 years old which can be interpreted as lunar cycle calendars. They may have been constructed by hunters who recorded each day and each week with a specific marker to indicate when the moon was full and when it was not. In that way, the hunters could keep track of how far and for how long they had been traveling when they returned home. Memory records such as these can then serve as the basis for the reconstruction of our notion of time.

Andreassen: This seems to be quite telling. Today, we also use calendars for planning, and they play a very important role in everyday lives. It might seem banal, but I guess this is an example of the power of external symbols and of how consciousness executes its governing powers. We are, of course, conscious when we make a note of what we are supposed to do sometime next week. Then, we forget it. However, returning to the entry gives an input to the brain, and thereby the action is initiated by a decision made in the past.

Donald: The timeframe for planning our lives is now very much externalized because we are so poor at recalling such details. In modern life, this is crucial for survival. The interactions between symbols and the conscious individual mind track the unfolding sequence of our lives and structure our experience of time. Accurate planning and the actual recall of past behavior are relying more and more on external devices in our very complex world.

In a sense, external symbolization is very unnatural. It has accelerated human life. In the Stone Age, life did not move at such a fast pace because the symbolic framework of life was thin and ephemeral. For the most part, lives were lived at a much slower and natural pace and closer to nature. The main indices of time were the rising and the setting of the sun, or the changing of the seasons, very natural things. The high-tech, high-pace society we live in is overwhelmingly determined by complex symbolic structures that reside in such things as atomic clocks, legal codices, technological networks, and negotiated treatises, all external to the brain. External symbols have played an extensive role in the regulation of public life. At one point in the Roman Empire, the calendar had 180 festival days. Half of the year was taken up by special feasts, all of which had significance to the citizens. This required the use of complex external memory devices. In modern society, we also use external symbols to regulate life: In the working world, time is scheduled to the minute under external symbolic control.

Andreassen: In view of these considerations, is it possible to conclude that consciousness or the capacity for awareness beyond the immediate timeframe gets its powers of governance only by externalizing its representations?

Donald: No, I would not say that. The power of consciousness is also enhanced by mimetic means. Do not forget that many social rituals and customs impose unconscious conformity with a given society. People are born within a particular cultural framework, and individuals can passively follow directions that have been set out deliberately and consciously in that society. This is important: Creating or altering social customs (for instance, inventing a calendar) is a supremely conscious process. However, the temporal routines of a thousand people can become a ritualized, automatized process that does not require any individual reflection. Thus,

automatization of behavior can take place in collectivities just as it does in individuals and is not always dependent on external symbols or unconscious decisions once it is established. Consciousness is the creator and regulator, not the day-to-day executor.

Andreassen: Is calendar regulation an example on how self-awareness is entrenched in very old cultural habits that are implicit in their influence on cognition?

Donald: To a degree. It is liberating not to have to use our awareness all the time. There are times when we just want to move along passively with the social conventions while focusing on personal agendas. But, at other times, the conscious mind must intervene and restructure those conventions. It is during such times that the tremendous power of conscious metacognitive regulation is revealed.

References

- Donald, M. (2001). *A Mind So Rare: The evolution of human consciousness*. New York: W. W. Norton & Co.
- Ledoux, J. (1996). *The Emotional Brain: The mysterious underpinnings of emotional life*. New York: Simon & Schuster.

Paper received April 2004; revised February 2007

Kristian Tylén

When Agents Become Expressive: A theory of semiotic agency

In this paper I will outline a theory of agency proposed by Alan M. Leslie. It focuses on the cognitive mechanisms underlying our immediate recognition of different types of causal agency. I will argue that a fourth kind of agency should be added to Leslie's list, i.e. semiotic agency. Semiotic agency designates our ability to recognize and interpret intended expressive behaviour and objects in our surroundings.

Whenever we use language we perform an intentional act of a causal nature. By producing signs we induce a change in the attentional orientation of the addressee. This mental change is an event, and the addresser is responsible for the event, i.e. is recognized as the causing agent. Thus, communication is agency.

In the following I will introduce the notion of *semiotic agency* focusing on our ability to recognize communicative intentions in the behaviour of other agents and symbolic artifacts – an approach to the act of enunciation that will likely be seen as a challenge to the prevailing use of the notion causation and agency.

Semiotic intention and attention

Sometimes a rock is just a rock. We may or may not notice it there on the ground. We may be annoyed by its presence just ahead of the lawnmower or admire it for its aesthetic 'rocky' properties and pick it up for our rock collection. Sometimes a rock is a tool. We may consider the rock useful in some instrumental context (breaking windows

or building fences, etc.). When we recognize a rock as a useful instrument for some activity or purpose, it is because it fits an internal intentional program. Our intentional attitude makes the cognitive attentional system ‘transform’ the surrounding objects into potential helpers (or harmers) of the project. If the project is a window that needs to be broken or a fence that needs to be built suitable rocks will suddenly constitute the foreground of our visual attention.

But sometimes the rock is a message. We recognize the rock as referring to something other than itself. Take a gravestone. It ‘tells’ us that someone died and is buried there. Perhaps it ‘tells’ us to honour this person’s memory. This is puzzling. How can a rock possibly tell us anything at all? What is it that suddenly makes the rock *communicatively significant* to us?

When we experience the rock as a message, it cannot be a product of our own intentional attitude toward rocks alone. It is not a ‘top-down attention’ in the sense that we are explicitly searching for ‘message rocks’ as could be said in the case with the ‘instrument rock’¹. In some fantastic way, the rock addresses us (bottom-up attention). Or more precisely – we are able to recognize the addressing intention of someone else in the rock. We understand that some cognitive being has acted upon this rock in order for it to represent something totally different from itself. And this immediate understanding detaches our attention from the rock as such and directs it to the content of the message.

As I would prefer to avoid any hocus-pocus supernatural explanation of this information exchange, I suppose that the communicative intentional action of the addresser left a recognizable trace on the rock. It could be some kind of manipulative modification of the surface of the stone like an inscription, a modification of the immediate surroundings of the rock with cut flowers, etc., or a simple displacement.

Another example: We often go around lifting our eyebrows. It doesn’t mean anything. But sometimes a lifted eyebrow means a lot; it is intended as a message and is somehow immediately understood as such. Generally it seems that what would appear to be the smallest, most insignificant displacement, transformation or manipulation of an object can make it a medium for communication and is immediately interpreted as so by the addressee. Verbal language is made out of such micro-differentiations in the phonology of speech and the corresponding graphic representations in written

¹ About top-down and bottom up attention see e.g. Oakley (2003).

language, even though this ontology does not solve the mystery. Our everyday life leaves dozens of traces in the surroundings whereas only a minor part is intended and recognized as communication. Still, we manage to sort out the immense amount of background noise and focus attention on the semiotically significant details of the surroundings.

Thus, to sum up the first problem, it would seem that we are somehow cognitively equipped to recognize human semiotic intention and that our attentional system is extremely sensitive to semiotic acts and expressions in preference to other kinds of world phenomena. However, this complicated problem already breeds a lot of related questions. For instance, what is a semiotic act, and how does it differ from other kinds of human acts?

A Theory of Agency

In his article “A Theory of Agency” (1993) Alan M. Leslie proposes a tri-partite theory of agency, with a cognitive and developmental psychological approach. His theory may prove valuable in this context, as it concerns our ability to distinguish a set of properties that differentiates agents from other physical objects and allows us to track and interpret different aspects of agency in our surroundings.

To Leslie, agency is understood as ‘*the core constraints that organize our early learning about the behaviour of agents*’ (Leslie 1993: 1). The notion of ‘core’ constraint indicates a modular, biological approach to the problem of agency. Thus, Leslie believes that, as a result of adaptive evolution, humans have developed a highly specialized information processing system that provides the basis for learning and understanding the behaviour of agents. The result is a ‘*sophisticated capacity to explain, predict and interpret their [the agents] behaviour*’ that has adaptive advantages in a socio-physical world (1993: 1).

Though we could already at this point contest Leslie’s theoretical grounding, his work makes a suitable platform for further investigations into the notion of agency. In the following I shall give a brief overview of Leslie’s theory, skipping most details and focusing especially on the last part, which unfortunately is the least elaborated and documented.

Leslie’s experimental observations of children from the age of 6 – 18 months motivates a typology of causality, hierarchically ordered in *mechanical causality*, *intentional causality* and *psychological causality*, in which each of these components corresponds to a

set of properties that characterize agents and distinguish them from other kinds of physical objects. These are as follows:

Agents have *mechanical* properties. They have an internal and renewable source of energy or force, i.e. they do not need to rely on external sources. They are 'self-propelled' and capable of bringing about changes in physical space.

Agents have *actional properties*. Agents act intentionally, in pursuit of goals and re-act to the environment as a result of perceiving. Furthermore, the acting and re-acting agents can get together and *inter-act*.

Agents have *cognitive properties*. The behaviour of agents is determined by internal cognitive properties, e.g. holding a certain attitude to the truth of a proposition.

(Slightly adapted version of Leslie 1993: 2, italics added)

While Leslie is most concerned about the first category and the physical force representation (and this is where his theory is most thoroughly documented) I will concentrate on the latter two.

Mechanical agency

During the first eighteen months of life, children gradually acquire the ability to recognize these different aspects of agency. From about six months, children show sensitivity to mechanical causality and the ability to distinguish internal-force bearing, self-propelled agents from mere physical objects. Central to the understanding of mechanical causation is the transmission of force, e.g. from an agent to an object, through a spatial (and temporal) contiguity. Without physical contact, no mechanical causal influence is possible.

Actional agency

Within the second half of the first year, children begin to recognize intention in the acts of other individuals, i.e. an agent acts in pursuit of an internal goal or reacts to an external aspect of the environment through perception. Furthermore, two or more

goal-pursuing agents can join in different kinds of cooperative or competitive interaction, such as helping or harming, depending on the mesh of goals. In this period of development, children also begin to follow the eye gaze of other people. Thereby the child will seek the reason for the agent's behaviour in the focus of her attention and action. Thus, the agent is now approached not only as a transmitter of force (in the physical/mechanical sense) but also as a possessor, a transmitter and a recipient of *information* (Leslie 1994: 143).

An important aspect of actional agency in opposition to physical/mechanical agency is the spatial and temporal detachment of, or 'distance' between, cause and effect. Thus, we don't find the same transmission of force through direct contiguity. When we witness an intentional goal-directed action of an agent, the effect or goal-state has not yet come about. It is still a little ahead in time, and may in fact never be realized at all. Likewise when we react to some perceived occurrence in the surrounding environment, this 'cause' may be spatially distanced from the perceiving and reacting subject. As this non-contiguous relation between cause and effect is only enabled by the interference of a cognizing mind, perhaps it would be more precise to follow Østergaard (2000) and term it *mental causation*². As Østergaard stresses, the difference in ontology between mental cause and physical effect makes it impossible to talk about a 'transmission of force' in the traditional sense (Østergaard 2000: 10). Cautiously, however, one could propose the notion of *pregnance* in René Thom's sense of the word (Thom 1990) as a kind of force underlying actional agency: Whenever a perceived situation leads to a reaction of an agent, it is because some element has a potential to influence the agent. Likewise, when an agent out of internal motivation directs her intentional actions towards a certain goal, e.g. the acquisition of an object, it is because the mental visualization of the goal-state represents a pregnancy to the agent (in this case an attraction towards conjunction with the object) (Østergaard 1998: chap. 1)³. This 'goal-status,' with its potential influence, is not (necessarily) a property of the objective physical world but is due to the biological and cognitive *values* of the agent. In a well-elaborated attempt to naturalize human intentionality, Gallese and Metzinger

² In his article "Mental Causation" (2000), Østergaard gives a detailed description of the mental processes intervening between a perceived physical event and a corresponding physical reaction from the observer-agent. These processes include perception, memory, belief, pro-attitude and motor control.

³ This is a very superficial presentation of an otherwise very elaborate theory; a more detailed presentation is beyond the scope of the present article. I should add, though, that René Thom distinguishes between internal and external pregnancies, that more or less correspond to my notions of biological and cognitive values (see e.g. Østergaard 1998: 15-33)

(2003) describe the tight neural connections between motor action schemes (e.g. the so-called F5 area of premotor cortex), the goal-representational system and the value system. Their conclusion is that we seem to possess certain ‘goal-related’ neurons. When confronted with a pregnant object, this system immediately triggers action simulations, i.e. motor schemes that correspond to the most rewarding or satisfactory behaviour towards the specific object. Interestingly, a set of these neurons seemingly do not mind the special perspective of action goal representation but are equally active in first-person goal-pursuing action and the third-person observation of other subjects directing their intentional behaviour towards an external object. These are the so-called ‘mirror neurons’, which would seem to play an important role in our understanding of others’ behaviour, thereby making it possible to adjust our behaviour to our social context. These neurons probably constitute the base of social cognition (Gallese and Metzinger 2003: 29)⁴.

Cognitive agency

From about the age of eighteen months, children gradually gain the ability to ‘detect’ volition, belief and pretence in agents. That is, they begin to understand the cognitive properties of other subjects. According to Leslie, this requires a sensibility to the relationship between agents and information (Leslie 1993: 11). Attending solely to the ‘novel meaning’ of the behaviour of an agent, the child becomes aware of the propositional attitude of the agent. She can understand some actions of the agent as caused by representational, fictional circumstances. Thus, when the agent asks the child to say goodnight to a teddy bear (my example) it is not because she misinterprets the actual situation, believing the toy bear to be cognizant of verbal language. The agent is not delusional. The child recognizes the intentional behaviour of the agent as related to (i.e. ‘caused’ by) an imaginative reality in which the bear is animate and cognitively equipped. To do that, the child must not only direct its attention to the actual referent of the behavioural expression, but likewise to the agent’s own attitude to the behaviour. In other words, the child must pay special attention to the way it is performed, i.e. intonations, special gestures or the facial expression of the agent that signifies fictional attitude. Furthermore, the child must itself be capable of simultaneously holding and

⁴ Other important brain areas in this respect are the so-called ‘Theory of Mind’ areas described by Amodio and Frith (2006).

keeping track of two representational inputs; one constituting the actual scenario in which the teddy bear is just an inanimate, ‘dead’ object, and another featuring the content of the agent’s behaviour, i.e. an animate and cognitively equipped bear⁵ (DeLoache 2004).

To Leslie, the notion of ‘cognitive agency’ only designates our ability to understand the behaviour of an agent as caused by fictive (non-actual) circumstances. Principally he tries to keep the third-person perspective on the agent, having the child on the sideline as an observer, interpreter and recognizer of cognitive agency. This project succeeds in some of his false-belief tasks where a child is asked to predict the behaviour of another subject in a classic ‘hidden object task’ (Leslie 2000). Nevertheless, most of his examples contain a lot more. Take his favourite example, repeated in all his writings on the subject (e.g. Leslie 1987: 417, 1993: 12, 1994: 141, 2000: 1236): A mother hands the child a banana saying, “The telephone is ringing. It’s for you”. As in my own teddy bear example, the child is not only expected to understand the behaviour of the adult agent as caused by a cognitive attitude, but is invited to take part in the pretence, that is, to attend to the specific *content* of the pretence and *act* accordingly. Most eighteen-month-old children will readily adapt the propositional attitude of the adult agent and follow it into the representational world of banana telephones. In the same way that the child in the earlier stage of development followed the eye gaze of the adult in actual space, it now understands the ‘invitation’ of the adult to follow her attention into a mental, representational space. But how does this come about? The recognition of pretence does not in itself explain the causal structure of the representational and participatory effect in the child. Leslie wilyly escapes this obvious problem of causality and agency. Though he mentions ‘sensitivity to information’ and even ‘communication’ as elements of cognitive agency, he goes to great lengths to leave out remarks on language. This is probably due to Leslie’s conviction that language forms a discrete and isolated cognitive module in itself and thus cannot be directly connected to the cognitive mechanisms of causality. Only in a couple of lines in his 1994 article (p. 142) does he mention language explicitly, confirming his belief in separate processes. However, this strict separation of language and communication must strike one as counterintuitive when we focus on the causal structures of semiotic exchange. How does an agent cause a subject to create mental

⁵ The act of pretending thus yields a capacity for holding dual representations, not losing track of the specific labels of the inputs: what is real and what is fictive (cf. DeLoache 2004: 4).

representations of non-present realities and afterward to attend to certain aspects of these? Trying to account for this intriguing problem, I will suggest a supplementary notion to Leslie's typology of agency, i.e. *Semiotic Agency*.

Semiotic agency

One could wonder why I chose to spend several pages of this paper lining up an elaborate theory of agency that does not once mention language and thus proves insufficient in solving our initial problem. Nevertheless, I approach the communicative act as a type of causation, which at the same time makes the question of agency relevant. Furthermore it might turn out that these 'information processing systems' are not at all as autonomous as Leslie assumes. The introduction of semiotic agency at least yields tight cognitive connections and 'overlaps' with the cognitive mechanisms underlying our recognition of other kinds of causality and agency, e.g. Leslie's notions of *actional* and *cognitive* agency.

As briefly noted in the prior sections, our ability to recognize actional and cognitive agency constitutes the foundation for the development of *social cognitive skills*. Some of these we share with our fellow primates, but at some crucial point in evolution something happened that made the species of homo sapiens develop in a unique way. Not hesitating to avoid the still problematic cognitive and neural details, this species-unique aspect of social cognition can be phrased more or less like an understanding that "*I am an intentional and cognitive 'self' among other intentional and cognitive 'selves' like me*" (Tomasello 1999: 4). This understanding enables us to identify with other subjects, their intentional and mental states – something that dogs, squirrels and hyenas probably do not do. Thereby it becomes possible for us to exchange and share experience and knowledge, i.e. it enables an especially effective *cultural transmission* and *cumulative cultural evolution* (I return to these notions below). Only humans seem to possess this ability that is the primary reason for our high degree of evolutionary adaptation (ibid.).

One of the most interesting and important aspects of social cognition is *joint* and *shared attention*, which probably constitutes the core foundation of semiotic agency and exchange. As noted in an earlier section, from early stages of ontogenesis, children tend to follow the gaze direction of others. But only at the age of nine to twelve months will they begin to recognize an intentional communicative use of attentional orientation; sounds and gestures that previously were only comprehended as dyadic emotional

exchanges are now understood as a volitional attempt to manipulate the attention of the child with respect to external objects (Tomasello 2000: 64).

The child understands the special behaviour of the adult as an invitation to participate in an intentional act. Not as an agent directing its own intentional acts towards a pregnant object, as when she is reaching out for a toy. Nor as a direct object of some other agent's intentional acting, as in the act of changing diapers, for instance.

The recognition of semiotic agency thus triggers a special governing cognitive schema (Brandt 2000: 4)⁶: A syntactic structure constituting a constrained set of actantial roles; first-person subject (addresser), a second person dative-object (addressee) and a third-person direct object (content referent). A basic property of this schema is the 'directedness' of the act, going from the first-person agent via the second-person patient and further on to the third-person object of attention. It can be phrased simply: "I want you to attend to X [= the object of joint attention]".

The structural properties of the schema can be illustrated with a simple example. Say I want to direct the attention of a child to a specific point in space, e.g. to share the experience of a passing dog in the street. Then one of the ways to go about it is to establish eye contact with the child and point to the dog using my arm and index finger. If my intention is recognized, the child will redirect her attention away from my eyes, along the imaginative line extending from my finger and aiming at the dog. Declarative pointing is in fact very illustrative in explaining the proposed structure of semiosis, as it features a decomposition of the communicative intention and the content specification of the message in two discrete acts: The eye gaze of the addresser can be interpreted as signifying: "This is a message! I want you to attend to this [the object pointed to]", while the act of pointing does the work of filling in the brackets by naming the referent content – the dog.

Comprehension of the directionality of communicational acts is crucial to the acquisition of intentional semiotic behaviour. In early pre-linguistic language acquisition, the child must be able not only to recognize the intentional state of an adult-agent with respect to a third-person goal-object, but also to recognize it as directed toward her own attentional states. This requires that the child is able to monitor herself as another intentional agent intended to engage in a specific 'role' of a

⁶ Brandt (2000) suggests that this schema is innately determined. I am a little more hesitant in that respect, but it is a fact that even very young infants are a lot better at grasping the basic "architecture of interaction" than any of our fellow primates (see for instance Tomasello *et al.* (1997).

joint activity. When the child comprehends the intentional state of adult as an invitation to share attention, it will follow the adult agent's attentional direction with respect to the third object.

The next step is to understand this fixed set of roles as interchangeable. Thus, to fully acquire semiotic behaviour, the child must engage in role reversal imitation. Again, this is not a simple task. The child cannot merely substitute herself for the adult with respect to a third-person goal-object as in other kinds of intentional doings. In the act of communication, the primary intentional goal is actually the attentional state of the second-person addressee. The child must thus also substitute the adult for herself in the role of second-person dative object, i.e. she must understand that the act of signification is directed towards the second-person and not directly at the reference object (Tomasello 1999: 103).

The triadic structure of semiotic exchange is what distinguishes it from the other object-object and subject-object oriented types of agency and is maybe the best argument for the establishment of semiotic agency as a discrete type.

Yet another constraining property of semiotic agency is its spatial structure. In opposition to the spatial structure of actional (mental) and cognitive causality that would have either a mental cause and a physical effect (intentional acting) or a physical cause and a mental effect (re-acting), both the cause and effect of semiotic causality seem to have a mental ontology: The addresser's inclination to share an experience constitutes the cause, while the reorientation of the addressee's attention is the effect. But the causal connection can only be established through an intermediate chain of causal events in physical space. The very act of signification must be physical. All kinds of signifying gestures need to have a minimal physical ontology, i.e. they have to be perceptually traceable in physical space in order to reach the addressee (Sinha 1999: 11). On the other hand, any object or physical property can gain referential symbolic meaning if an agent intends it to do so and the recipient recognizes this intention. Intentionality is thus both a necessary and sufficient criteria for symbolicity (DeLoache 2004: 67). We can easily establish ad hoc symbols, agreeing that the spoon is a car or (to take up Leslie's example) that a banana is a telephone, in a specific game or situation. But like the case of declarative pointing and other communicative gestures treated above, the intention that 'this object or gesture is now to be understood symbolically' is

mediated through expressive behaviour such as gaze and facial expression⁷. It is the intentional *acting with the object* and not the object itself that makes it a symbol. And a moment later, the spoon will be a spoon again and the banana will be eaten.

While we still have all three participants of the communicative situation present in the same setting, with the addresser and the addressee facing each other, we can explain the signification and recognition of communicative intention by extra-linguistic conditions such as eye gaze, gesture etc. But much of our everyday communication does not proceed in these face-to-face intersubjective contexts. Throughout cultural history, oral communication and ad hoc symbolization has been supplemented with other modes of communicative exchange, using conventional cultural artifacts such as graphically represented symbolic sign systems, figurative and auditory media, aesthetic objects, etc. But when the sign does not any longer have the intentional, behavioural support, the accompanying gaze and gestures, what then constitutes the ‘addressing’ mechanism of symbolic artifacts? Or put differently: How is the communicative intention signified in such a way that the potential addressee knows that the object at hand is to be approached as a piece of symbolic communication? That she is supposed to take on the role of a dative-object in relation to the object? Let us first have a look at written language.

There is no doubt that the ‘invention’ of linguistic sign-vehicles has been of tremendous importance for human cultural evolution, as it made an especially effective accumulation and exchange of cultural knowledge possible. By means of ‘materializing’ the oral, symbolic gestures, communicational exchange can be detached from the moment of utterance, i.e. a message can be produced in the absence of the addressee and received in the absence of the addresser. The message is less transient: potentially numerous subjects can continuously consult it over time. But another important property of such conventional symbol systems is that they are artifacts, (almost) exclusively employed as a medium for interaction. Unlike the situation with the spoon-car and the banana-telephone, we are not dependent on the ‘acting out’ to know that they are intended symbolically. The symbol-objects are not of any particular interest in their own respect as they do not lend themselves to other instrumental activities (for instance they can’t be eaten). When fully acquired, linguistic signs are so cognitively

⁷ The recognition and interpretation of facial expression probably constitutes the most basic level of primate expressive behavior and we even seem to be equipped with special neural facilities to recognize and interpret facial expressions.

integrated it almost takes an academic degree to be able to disregard the referential aspects and appreciate their phonological or graphic object-like properties⁸.

In the process of conventionalization or entrenchment of a symbolic sign system, the sign-objects themselves seem to become ‘bearers’ of the communicative intentionality. Consequently the communicative intention of the addresser is *not* explicit and immediately distinguishable from the referential function, the ‘pointing’. They do not, it would seem, constitute discrete properties. Let us take the example of linguistic texts: Confronted with a piece of written verbal text, we will probably not doubt that it is a symbolically coded message, intentionally produced by someone for interactional purposes. We recognize this even before reading it. In fact, we are usually capable of recognizing intended linguistic communication even when we cannot ‘access’ the semantics of the specific sign system, e.g. a foreign language.

But far from all of our communicative uses of objects and artifacts are conventionalized in the way that verbal language is. Cultural convention as such cannot solve our problem. For instance, we readily recognize the addressing intention in works of conceptual art that do not in any way reproduce a cultural convention. If by coincidence we should come across Marco Evaristti’s red iceberg⁹ (not knowing that it was an artistic installation) we would probably spontaneously categorize such an unexpected variation of our surroundings as socially motivated and communicatively significant. This is due to a dynamic temporal reading of the seemingly static scenario. In other words, we understand it in terms of its causal history (Leyton 1992: 157). From the organization of the scene we infer the type of events, acts and agents that could have caused it. Naturally, these kinds of inferences are not one-to-one reconstructions of the past but rather work as a kind of natural probability calculation: What is the most likely causal process leading to the present state?

Non-expected variations of our surroundings can be caused by natural forces or by accidental or intentional instrumental human acts and each of these kinds of cause leave their more or less specific traces on the scenario. A set of such traces will motivate a communicative reading. That is, the scenario actualizes a previous intentional act that is directed at the manipulation of our attention. If we return to the

⁸ This is probably the main motivation when Andy Clark (1997) states that verbal language is “the ultimate artefact”. He hereby stresses the well-integratedness of an external symbol system in our cognitive processes and compares it to other kinds of (more instrumental) tool-manipulations.

⁹ See pictures of Marco Evaristti’s installation “The Ice Cube Project” at the URL: <http://www.evaristti.com/Work/performance/ICE/index.htm>

iceberg scenario, it tells the story of an agent giving at least some effort and concentration on (re)organizing the scene in a way that strikes us as purposeful. Icebergs are simply *not* red. Of course we could imagine some one spilling some paint or a dining polar bear leaving some traces of seals' blood. But a crucial fact about this particular scene is that the iceberg is entirely covered with red colour. The probability that this could have happened by chance is infinitesimally small. We thus have two principles guiding our intuition that this is intended by some agent to be a piece of aesthetic communication: 1) The bringing together of two elements, an iceberg and red colour, that are not normally part the same scenario, and 2) a certain 'well-orderedness' or symmetry in the way the elements are arranged or combined. Analyses of a series of other communicative scenarios reveal a similar pattern: The combination of the two principles seems to be of a general character¹⁰.

Communicative signs constituted by scenarios such as the Ice Cube Project do not gain symbolic meaning by bare recognition of cultural conventions. They do so by reference to the intentional causal history we are able to reconstruct from them. This is due to the interpretative nature of sign types. The symbolicity of a sign-object is not an intrinsic property of the object, but is dependent on the attitude of the interpreter. We can thus approach the same object with regard to its similarities with other objects (iconic reference), or its correlations with other things (indexical reference). Furthermore, we can approach the object with regard to its "*involvement in systems of conventional relationships*", i.e. as a symbol (Deacon, 1997: 71). Though a sign-object is intentionally used in only one of the referential functions, say symbolically, and interpreted as such, there might be intermediate interpretative processes in the reception of the sign. This is at least indicated by the structure of symbolic reference proposed by Terrence Deacon.

The hierarchical nature of symbolic reference

In his interesting work *The Symbolic Species* (1997), Terrence Deacon proposes a cognitive model of symbolic reference hierarchically based upon sign systems of icons and indices. Founded in evolutionary neurobiology, his point is that the enigmatic phenomenon of symbolic reference did not emerge out of nothing but can be

¹⁰ Here I refer to a work in progress related to my PhD thesis on communicative intentions and symbolic artifacts.

approached as a higher order extension of existing sign systems shared with our fellow primates. The building blocks for symbolicity were already present in earlier stages of our neural biological development. To roughly sketch Deacon's model, indices rely on relations among icons: Our immediate recognition of a perceptual phenomenon, say the smell of smoke, is due to its iconic relation to previous experiences of smoke. But it also tells us that something is burning. Former experiences of spatial contingency of fire and smoke create a stable correlation between the icons. Whenever confronted with one of them we will expect the co-presence of the other. Smoke means fire. Central to the conditional nature of indexical reference is thus the continuous spatial or temporal contiguity of sign and reference. Should we suddenly begin to have experiences of smoke that are not accompanied by fire, the reference will soon break down (Deacon 1997: 82). Likewise, when we teach our dog to respond to linguistic signs like "food", "sit" or "time for a walk" it will simply understand the correlation of that phonological string and some activity or object. The reference is indexical (though we may expect it to be symbolic) and can only be maintained through a continuous correlation. That is the funny thing about symbols: They do not depend on contiguity of sign and reference. We can easily point to objects that we do not have any prior experiences of, or which do not exist, such as dragons, UFO's and teenage mutant ninja turtles. Many children have words like "tiger" and "witch" in their early vocabulary, despite the fact that they have not had any (non-fictional) experiences of these phenomena.

Deacon explains this problem by the special relational character of a single symbolic reference in relation to a system of symbols. Symbolic reference is thus formed by systematic relationships between index tokens (that again are formed by icons). When a logical combinatorial system of relationships among indexical tokens is stabilized, we begin to rely on the relation of tokens and the contiguity of sign and object is no longer necessary for the maintenance of reference (Deacon 1997: 87)¹¹.

¹¹ In the early stages of language acquisition the child learns words for things already familiar. In some sense, the process goes from perceived object to symbol. But gradually, as the relational system of symbols is adapted, the symbolization process can be inverted. It now goes from word to object. The child will learn new words for which the reference still needs to be stabilized. The reference object may be less familiar or even unknown to the child. Often she will temporarily mix up the novel word's reference object (e.g. mistake 'lion' for 'tiger' etc.).

The index of semiotic agency

The hierarchical dependency of sign types in symbolic representation may help to explain our initial problem concerning the signification of semiotic agency, though we may end up somewhere else than Deacon intended. When we are confronted with an intended symbolic sign-object, our first and immediate cognitive routine is probably one of recognition. We recognize it as a symbolic sign because of its similarities with previously experienced symbolic signs. This basic iconic reading precedes the reference aspect. But the recognition of the special ‘sign features’ of the object changes our attentional attitude toward the object. We no longer perceive the object as such, but relate it to its causal history; the act that created or manipulated it. We thus experience the object or scenario as a kind of metonymical extension of the intentional agent responsible for it. In this sense the sign-object is primarily indexical; the object points to its cause: The communicative intention of an agent. The triadic actantial schema is triggered and we become dative-objects. Only then does the question of symbolic reference become relevant – it is secondary to the indexical comprehension of semiotic agency.

Now the remaining question is of course the nature of ‘sign features’, that is, the perceivable features of an object that tell us that some agent has acted upon it in order to make it stand for something else, to signify something other than itself.

Though some of the structural properties of semiotic agency may in fact be neurobiologically hardwired, Tomasello stresses the importance of *cultural transmission* and *learning* in the acquisition and understanding of symbolic sign systems. The evolution of the sign-symbol is thus probably not to be traced in the neural architecture of the individual, but in the shared structures of cultural knowledge and cognition (Tomasello 1999: 11). Symbolic sign-objects are cultural artifacts. They cannot be compared with the ways our fellow primates communicate using smells and the like. These behaviours are innate and have evolved in evolutionary time. Thus, symbolic sign-objects are products of an accumulative cultural process: By a gradual and steady stabilization of conventional object-meaning pairings, handed over and further developed from generation to generation, symbolic language finds its present form.

In a socio-cultural world, survival is to a large degree dependent on one’s ability to get on in symbolic exchanges. From the earliest stages of ontogenesis children get an intuitive understanding of symbolic signs as valuable objects. Through a kind of

‘conditioning process,’ symbolic signs are highly invested with pregnancy. In acquiring language, the child is continuously rewarded by an increasing adaptation to its social surroundings. It is satisfactory for the child to understand and be understood, and, most importantly, to be able to participate in ever new social practices.

The general human attentional sensitivity to semiotic gesture is probably due to these kinds of socio-cultural adaptive processes. To reach a high level of adaptation, we have to attend to meaning by differentiating micro-details in the phonological and graphic behaviours that constitute language. We gradually acquire an attentional sensitivity towards a broad range of conventionalized sign-systems, from the ‘language’ of traffic lights and aesthetic artifacts to culturally determined behavioural practices, dress codes and complex grammatical constructions.

It might be that our experience and increasing familiarity with various conventional symbol systems motivates a more abstract understanding of ways that objects can be manipulated for symbolic purposes. We begin to recognize certain kinds of acts and object manipulations as communicatively intended, not because they reproduce a convention but because they reproduce a more abstract behavioural pattern. The bringing together of objects and new contexts, combined with the principle of well-orderedness, seems to be a powerful way of expressing oneself. In the examples treated above, it is the unconventional combination of objects, properties and context that motivates the indexical reading of communicative intentionality. It is very unlikely that these different elements would be brought together by natural forces or accidental acts. They are most probably the product of an intentional expressive act.

The same goes for our initial problem of the rock. Even though we might not be familiar with the specific conventional uses of rocks (on graves, mountain peaks, etc.), we are probably still sensitized to the perception of rocks, standing in an upright position against all probability calculations of its ‘natural tendencies’ (forces of nature tend to have oblong rocks lying down). The static rock scenario is thus interpreted in a dynamic fashion as part of the causal history. And the well-ordered, non-natural and non-coincidental characteristics of the scenario point to the intervention of human intentionality. We recognize the otherwise superfluous human act of erecting rocks as expressive and significant, as an index of semiotic agency¹².

¹² It would seem that instrumentally superfluous acts are especially suited for communicative use. Banging hammers against walls, for instance, would not qualify as a good way of creating a language, as the communicational use would be hard to separate from the instrumental. We would have to cancel all kinds of

It would seem that the borders of intersubjectivity extend far into the physical domain. What at first glance seem to be static physical states and properties of our surroundings are experienced as highly dynamic processes of interpersonal communication. Almost any variation in physical space (or even lack of expected variation) can be interpreted as socially grounded depending on the context. The ‘variation’ thus bears the trace of a socially motivated intentional act, which changes our attitude to the scenario – we take the symbolic stance.

Knowing the specific cultural conventional use of rocks on graves we can thus access the intended reference of the symbol. But the recognition of a communicative intentional act is always prior to the decoding of referential meaning.

References:

- Amodio, M.A. & Frith, C.D. (2006). ‘Meeting of Minds: the medial frontal cortex and social cognition’. *Nature Reviews, Neuroscience*, vol. 7: 268-77.
- Brandt, P.Aa. (2004). ‘From gesture to theatricality’; in P.Aa. Brandt. *Spaces, Domains, and Meaning*. Bern: Peter Lang Verlag.
- Clark, A. (1997). *Being There. Putting Brain, Body, and World Together Again*. Cambridge, MA: MIT Press.
- Deacon, T.W. (1997). *The Symbolic Species – The Co-evolution of Language and the Brain*. New York: W.W. Norton.
- DeLoache, J.S. (2004). ‘Becoming symbol-minded’. *Trends in Cognitive Sciences*, 8 (2): 66-70.
- Gallese, V., & Metzinger, T. (2003). ‘Motor ontology: the representational reality of goals, actions and selves’. *Philosophical Psychology* 16 (3): 365-388.
- Leslie, A.M. (1987). ‘Pretense and representation: The origin of “theory of mind”’. *Psychological Review*, 94 (4): 412-426.
- Leslie, A.M. (1993). ‘A theory of agency’. *Technical Reports of the Rutgers University*. London: Center for Cognitive Science.
- Leslie, A.M. (1994). ‘ToMM, ToBy, and Agency: Core architecture and domain specificity’; in L.A. Hirschfeld & S.G. Gelman (eds.). *Mapping the Mind: Domain specificity in cognition and culture* (pp.129-148). Cambridge: Cambridge University Press.
- Leslie, A.M. (2000). ‘“Theory of Mind” as a Mechanism of Selective Attention’; in M.S. Gazzaniga (ed.). *The New Cognitive Neurosciences, 2nd edition* (pp. 1235-1247). Cambridge, MA: MIT Press.
- Leyton, M. (1992). *Symmetry, Causality, Mind*. Cambridge, MA: MIT Press
- Oakley, T. (2003). *A Grammar of Attention. A Treatise on the Problem of Meaning*.
<http://www.cwru.edu/artsci/engl/oakley/papers.htm>
- Sinha, C. (1999). ‘Grounding, Mapping and Acts of Meaning’; in T. Janssen & G. Redeker (eds.). *Cognitive Linguistics: Foundations, Scope and Methodology* (pp. 223-255). Berlin & New York: Mouton de Gruyter.
- Talmy, L. (2000a). *Toward a Cognitive Semantics*. Vol. I: *Concept Structuring Systems*. Cambridge, MA: MIT Press.

instrumental banging. Small finger rings, big bronze statues, silly dance steps and strange oral behaviors etc., on the other hand, are excellent for expressive use as we could not imagine them employed in any other sensible connections.

- Talmy, L. (2000b). *Toward a Cognitive Semantics*. Vol. II: *Typology and Process in Concept Structuring*. Cambridge, MA: MIT Press.
- Thom, R. (1990). *Semio Physics: A sketch*. US: Addison-Wesley.
- Tomasello, M., Call, J. & Gluckman, A. (1997). 'Comprehension of Novel Communicative Signs by Apes and Human Children'. *Child Development*, vol. 68 (6): 1067-80
- Tomasello, M. (1999). *The Cultural Origin of Human Cognition*. Cambridge, MA: Harvard University Press.
- Tomasello, M. (2000). 'First steps towards a usage-based theory of language acquisition'. *Cognitive Linguistics*, 11 (1-2): 61-82.
- Østergaard, S. (2000). 'Mental Causation'. Århus: Center for Semiotics, University of Aarhus.
- Østergaard, S. (1998). *Kognition og katastrofer – studier i dynamisk semiotik*. Copenhagen: Basilisk.

Paper received April 2004; revised February 2007

Agents without Agency?

The notion of agency occupies a central position in several of the cognitive sciences, particularly artificial intelligence (AI) and robotics. However, the notion largely rests on folk psychology and is usually left more or less undefined. This paper examines different notions of agency and analyzes the concept with a special focus on its role in AI, where much recent research has been devoted to the construction of artificial agents. We highlight recent naturalist theories of agency and argue that even if agency is not merely a folk-psychological concept without ontological bearing or scientific value, the phenomenon is more complex than most theories acknowledge. We argue that, as the title implies, none of the so-called agents of contemporary AI and robotics can be attributed agency in the strong sense, although artificial agency is not impossible in principle.

1. Introduction: Artificial agents in the complexity paradigm

The ambition to synthesize intelligent autonomous agents using different types of information technology, as pursued for fifty years by a growing number of research fields (from traditional AI over robotics to software agents and artificial life) remains without significant success, at least when it comes to human-like intelligence, and especially when viewed in relation to the vast resources devoted to the enterprise. Reasons for this scientific misfortune are as manifold as they are varied. Generally the

quest for AI¹ is severely hampered by its philosophical load; creating intelligent artifacts simply requires wrestling very tough philosophical questions on the nature of agency, autonomy, cognition, etc. These are quite elusive and intricate concepts that still need clarification, not only within AI research. Agreeing on a terminology and a common understanding of the slippery concepts involved has so far been without success, even though the issue has been given considerable attention. Consequently, the notions of agency and autonomy hardly enjoy any common understanding despite their pivotal role in AI. Many AI researchers simply stipulate their own version of these concepts, often closely related to the specific nature of the artificial agents they are trying to create. All in all, even if technical obstacles pose considerable challenges for AI, especially with an increased focus on the morphological and material aspects of intelligent systems, the primary battlefield remains conceptual. This paper aims at contributing to the effort of clarifying and qualifying the notion of agency while acknowledging that success lies in continuous analysis and discussion, not in final definitions.

However, as for any complex phenomenon, the confusion probably relates to the very effort of chasing ontological essences. Stuart Russell and Peter Norvig succinctly capture the spirit of a dominating pragmatism when they state: “The notion of an agent is meant to be a tool for analyzing systems, not an absolute characterization that divides the world into agents and non-agents” (Russell & Norvig 1995: 33). The contemporary naturalistic climate, understood as efforts toward a descriptively pluralistic but ontologically unified understanding of reality, generally opposes rationalistic, bio-chauvinistic or other metaphysically biased theories of agents and is certainly friendly towards a multifaceted and graduated understanding of agency – including the synthetic enterprise of AI, supposedly only feasible with a fairly liberal notion of agency.

On the other hand, acknowledging the inadequacy of ontological essentialism and the need for multiple complementary models to meet the complexity of the world is no invitation to a theoretical free lunch. Expanding the conceptual toolbox through cross-disciplinarity requires very careful deployment of terminology to avoid conceptual dilution and theoretical vacuity. So, although contemporary naturalism per principle allows for the synthesis of agents, the inherently processual and complex understanding

¹ If not otherwise specified we include robotics and other related areas dealing with how to synthesize agency when referring to AI as a generic field.

of cognitive phenomena dominant within the naturalistic paradigm has left the enterprise of creating agency immensely difficult.²

This paper describes the status of AI and discusses the prospects for synthesizing agency on the basis of relevant concepts, theories and tendencies. Section 2 characterizes the motivations behind the move toward complex and dynamic models for cognitive systems and explains why this tendency is especially crucial for understanding agency. In section 3, we provide an account of agency as conceived of by modern (interactivist, constructivist) models of cognition, which are characterized by a highly distributed and dynamic conception of agency in accordance with general theories of complex adaptive systems. In the second part of the paper, these models are compared to the broad use of the notion of agency within AI. Section 4 presents and analyzes notions of agency as deployed within AI, and in section 5, we assess the prospects of artificial agency on the basis on the characterization of agency offered in section 3.

2. Agents and agency

In endorsing a dynamical conception of the world, which has slowly grown dominant in the course of scientific history, it is sometimes helpful to insist on contrived distinctions between *processual* and *substantial* characteristics of phenomena. This scientific self-discipline serves to oppose an inherent tendency in human cognition and language towards reification.³

To motivate the departure from substantialism, as a blind alley cutting us off from explaining fundamental characteristics of agency, let us briefly sketch some problems with a substantialist conception of agency which inherits many of the difficulties of a dualist mind-body position. For instance, it cannot explain the *genesis* of agency, whether evolutionary or developmental (the problem is logically identical), since substantialism does not allow for the emergence of new qualities, let alone causal capacities (Bickhard 2000, 2003). Giving *things* ontological and causal priority to

² On a different note, assessing the deployment of, e.g., the term ‘agent’ within present AI research, we should acknowledge that the pursuit of artificial agency is subject to intense cultural interest and expectations. As such even serious AI research is always imperiled by powerful science fiction myths of intelligent robots inhabiting our world in the near future.

³ As witnessed by the history of ideas depicting a general move from anthropomorphic and animistic metaphysics over classical substantialism to a processual conception of reality growing dominant with modern physics and theories of complexity. In this context it is helpful to focus on *agency* instead of *agents* as a heuristic prosthesis to prevent slipping back into an infertile substantialism.

processes, you cannot explain how new entities arise from organizational processes (or in any other way), and you are stuck with the basic characteristics of your substances.⁴ This leaves us with different, but equally unattractive possibilities: Either the agent is a distinct entity (similar to the mind in the mind-body dualism) inhabiting a parallel ontological realm not abiding physical laws, or the agent is merely *identical* to physical processes, presumably in the nervous system and not a ‘real’ phenomenon of explanatory significance for a theory of organisms. (Our realistic leanings prevent us from considering the mentalistic possibility). The substantial stance on agency is also hard to maintain in the light of the vast array of different levels of agency manifested by different organisms. Having no account of its genesis or refinement, agency becomes a capacity a system either has or has not. Thus, either all animal life forms have the same ‘level of agency’ as human beings or only human beings possess agency. Either way, the notion of agency becomes counterintuitive.

A dynamic model, on the other hand, allows us to understand agency as a graded capacity correlated with the complexity of the organism in question both in relation to its evolutionary and developmental level. Agency thus becomes a measure for the amount of interactive possibilities an organism is capable of managing, and thereby its decoupling from immediate environmental possibilities. A tight correlation between environmental cue and reaction (popularly referred to as ‘instinct’) means less agency. Accordingly, a dynamic approach allow us to explain the evidently high ‘agency factor’ of humans without isolating mankind in nature or reducing other animals to mere automata. A fortiori this naturalistic, inherently processual and graded, notion of agency allows for synthetic kinds of genuine agents as pursued by AI research, in principle at least. Understanding and modeling the intricate organizational processes underlying agency in order to create them artificially is still a long way ahead.

2.1 Agency and entitativity

In accordance with the substantialist bias in classical thought, the agent has ontological priority over agency in most accounts and in common sense. However, if we do not buy into the substantialist conception of agents, the ontological and logical priority

⁴ A substance framework commits us to everything being either blends of the substances or structures of the basic atomic parts (depending on whether or not the substance framework is one of matter or of atoms) not as novelties emerging from organizational processes (Bickhard, personal communication. See Bickhard 2000 for an in-depth analysis).

cannot be such. In a nutshell, while agency serves as an *explanans* in the folk-psychological understanding and likewise in most classical philosophy, agency is the *explanandum* for processual naturalistic cognitive theories. Let us therefore investigate the relation between agents and agency in some more detail.

Donald Campbell provided a relevant account of reification mechanisms with the notion of *entitativity* (Campbell 1958, 1973). Entitativity refers to our notorious tendency to ‘entify’ phenomena proportionally to the number of coincident identification boundaries to the phenomenon. The more coincident boundaries and invariance through processes (i.e. through time) we are able to identify, the more robust entitativity or ‘thing-ness’ the phenomenon possesses: “Clusters of high mutual common-fate coefficients are our entities” (Campbell 1973). A diamond, one of the most robust and persistent things in our natural world, therefore possesses much entitativity. An organism possesses less, a soccer club might represent a borderline case, and a gas surely possesses no entitativity.

Our tendency to ‘entify’ phenomena is part of a group of cognitive mechanisms that probably operate statistically to reduce informational complexity and provide us with a rough, fast and adequately reliable overview of our surroundings: “A common preverbal evolutionary background strongly biases in the direction of finding “stable” “entities” highly talkable-about”, which are also prominent in language learning (Campbell 1973).

As we shall see shortly, such descriptive characteristics correlate neatly with the autonomy of cognitive systems and are not without taxonomic merit. Nevertheless, in relation to agency this categorical ‘rule of thumb’ easily facilitates erroneous conclusions. The reason is that the characteristic autonomy or organizational *cohesiveness* of cognitive systems is not identical to organizational *isolation* (cf., e.g. Varela 1997, Ziemke 2007). Since organisms are the center of many interactivities (they live, die, mate and interact as a unified system) and thereby the locus of a number of coincident organizational boundaries, it seems unproblematic to view organisms as possessing an (‘entified’) agent.⁵ Organisms are normally enclosed by a discrete physical boundary

⁵ In certain cases, agency determined on the basis of coincident boundaries has problems accounting fully for the observed phenomena. Social insects, for instance, cannot always be meaningfully demarcated individually as agents but only in their collective swarm behavior. Studies of social insects commonly point to the fact that communities such as anthills and termite nests are best described as unified agents due to the intricate interactive but cohesive organization of the individual animals’ behavior (e.g. Bonabeau *et al.* 1998). There are

and singled out by the fact that within this boundary disruption of organizational parts is much more severe than for external parts all things being equal.⁶

Still, that does not mean that agents are *organizationally* encapsulated within this entitativity. For instance, it is widely accepted that most western people today would do poorly without a lot of the tools civilization provides. Even on a less radical measure, we are heavily bound up with external cognitive “scaffolding” (Bickhard 1992, Clark 1997, Susi & Ziemke 2001) or “distributed cognition” (Hutchins 1995) as cognitive “cyborgs” (Clark 2003) doing ‘epistemic engineering’ (Sterelny 2004). So, granted that human agents are (at least apparently) the center of many processes of importance for their life they do not themselves encompass more than a fraction of the organizational and infrastructural processes they depend on. Instead agents exploit numerous invariants in the environment as cues enabling planning and real time negotiation rather than control in what has come to be known as ‘situated cognition’ (e.g. Clancey 1997, Ziemke 2002). Human agency is just one, albeit salient, organizational locus within a vastly complex hierarchical organizational web comprising processes ‘beyond’ the range of our control. For instance, we do generally only influence political matters in a minimal and highly non-linear way even though they might be of great consequences to our lives. If taken in the traditional rationalistic sense, human agency does not even control crucial bodily functions such as autonomic processes indispensable for surviving.⁷

That, on the other hand, does not mean that adaptive, interactive systems cannot be demarcated. But we should not look for physical markers. In theories of complex adaptive systems, processes that both contribute to the functional coherence of the system and depend on other processes of the system to function are to be considered constitutive parts of the system. One of the most prominent figures of this tradition, the Chilean neurobiologist Francisco Varela represented a very exclusive variant of this approach:

Autonomous systems are mechanistic (dynamic) systems defined as a unity by their organization. We shall say that autonomous systems are organizationally closed.

differences though, which should be noted. Agents are paradigmatically organisms, and the highly heterogeneous organization of organisms is different from the much more homogeneous “many-body” (Auyang 1998) organization of swarms.

⁶ Cutting off a lung patient from respiration aid technology is of course more severe than impairing the patient’s visual capacities.

⁷ The potential harm to human (normative) autonomy caused by this impotence is minimized by the dualist move of loosening the bonds to the flesh in classical rationalism.

That is, their organization is characterized by processes such that (1) the processes are related as a network, so that they recursively depend on each other in the generation and realization of the processes themselves, and (2) they constitute the system as a unity recognizable in the space (domain) in which the processes exist. (Varela 1979: 55)

Insisting on organizational closure as a criterion for autonomy, and consequently agency, Varela represents an extreme of the systemic understanding of agency as organizationally demarcated autonomous systems.⁸ But Varela's account still differs from the traditional entified notion of agency by being processual: The existence of the organism/agent is a continuous, recursive, and open-ended process of maintaining cohesiveness.

To summarize this section about the difference between agents and agency, autonomous systems participate in a host of processes on different organizational levels and with different degrees of influence. Hence, sticking to the notion of a causal monolithic 'agent' easily misleads into simplification and reification. It is more fruitful, albeit much more difficult, to deploy a processual approach focusing on the dynamic characteristics of 'agency' as an ongoing organizational capacity by self-maintaining systems.

3. Agency: From control to orchestration

History has repeatedly demonstrated scientific progress through the replacement of agency-based, substantialist and highly localized explanations by models of distributed complex processes (Browning & Myers 1998): Multiple animistic principles have been replaced by natural causal explanations; phlogiston was replaced by combustion, caloric with thermal heat, vital fluid with self-maintaining and self-reproducing organizations of processes; atoms are slowly giving way to quantum fields, and even the number of die-hard advocates for a reified Cartesian soul is decreasing (even though his explanatory dualism still lingers on, cf. Bickhard & Terveen 1995, Wheeler 1997). Generally speaking, the better science gets at tracking the organizational composition of

⁸ Some would argue that this brings Varela into trouble when trying to explain cognition, as cognition normally characterizes a *relation* between a cognitive system and its environment (cf. Bickhard, in preparation). However, see also Varela's (1997: 82) clarification that in the theory of autopoiesis the term *operational closure* "is used in its mathematical sense of recursivity and not in the sense of closedness or isolation from interaction, which would be, of course, nonsense" (cf. Ziemke 2007).

complex phenomena, the more references to discrete entities seem obsolete. This is especially true for the notion of agency itself. William Wimsatt writes:

It seems plausible to suggest that one of the main temptations for vitalistic and (more recently) anti-reductionist thinking in biology and psychology is due to this well-documented failure of functional systems to correspond to well-delineated and spatially compact physical systems [...] It is only too tempting to infer from the fact that functional organization does not correspond neatly to the most readily observable physical organization – the organization of physical objects – to the howling *non sequitur* that functional organization is not physical. (Wimsatt, forthcoming).

The case of consciousness provides a very relevant example. Despite the phenomenological unity of consciousness, modern neurological and cognitive studies depict consciousness as the result of an intricate synthesis of many contributing neural and distributed cognitive processes (cf. Edelman 1992). Consciousness may thus be a salient phenomenological phenomenon and also a fairly sharply demarcated subject of scientific description, but nevertheless not an ontological entity. Likewise, agency emerges from the integration of multiple distributed processes and does not represent a discrete source of action. So, even though ‘agent’ mostly refers to the capacity of causal initiation and spontaneity,⁹ agency is much about the *orchestration* of energy flow. Accordingly, the related notion of autonomy does not mean ‘self-governing’ in any isolational sense, but merely indicates a certain degree of dynamic decoupling from the environment allowing the system to promote its own interests. Autonomy is measured by the degree to which the rules (or more correctly norms, see below) directing interactive processes between a system and its environment are created and/or governed by the system.¹⁰ In fact, self-governing systems are always open systems (otherwise they might not even need governance) and open systems are intrinsically dependent on interaction with their surroundings. Autonomous agency can therefore only be understood as a system’s self-serving interactive organization of energy flow and not as a primal *causa efficiens*.

⁹ Agent comes from ‘*agens*’ denoting the driving (acting) part of a causal relation whereas ‘*patiens*’ is the passive, receiving part (Merriam-Webster Online Dictionary).

¹⁰ As the term ‘interactive’ indicates, the rules governing autonomous systems are constrained by various structures involved in the interactions and cannot be created uninhibitedly (as some radical idealist theories erroneously suggest). This notion of autonomy is closer to the political meaning than the philosophical (metaphysical) one (Christensen & Bickhard 2002).

3.1 Agency as self-organization in complex systems

Recent naturalistic models of cognitive agency place cognition in the high end of a continuous spectrum of self-organizing systems of increasing complexity. Under notions such as autonomy or self-maintenance, agency is linked to the very dynamic organization of adaptive systems in what Peter Godfrey-Smith calls “strong continuity” between models for life and cognition (Godfrey-Smith 1998, Wheeler 1997, Ziemke & Sharkey 2001, Ziemke 2007). Agency is explained as an emergent phenomenon arising from the self-maintaining dynamics in complex adaptive systems. In a circular manner, agency serves to further improve the self-maintaining processes by providing better integration and coordination of diverse system needs and interactive possibilities.¹¹ The continuity models share a primarily systemic approach to agency, stressing the overall viability (functional cohesiveness) of adaptive systems as the primary constraint for cognition. But as an inherently dynamic approach, historical factors also play an important role as non-trivial ways for past states of the system to influence subsequent behavior.

The systemic approach to agency has a long history but has left the marginalized outfields and slowly approached the center of the theoretical landscape of naturalist cognitive theories during the last decades. The epistemologies of the American pragmatists William James, John Dewey and Charles Sanders Peirce were congenial forerunners for the systemic approach as was the constructivism of the developmental psychologist Jean Piaget. Within theoretical biology especially the work of Jakob von Uexküll, the pioneer of ethology and bio-semiotics, is commonly mentioned as an early example of the linkage between the structural organization and cognition of organisms (e.g. Uexküll 2001; cf. Sørensen 2002a, Ziemke 2001, Ziemke & Sharkey 2001). Uexküll used descriptions of simple animals to stress the correlation between fundamental bodily needs and sensory-cognitive capacities. The classical example is the interactive environment (“Umwelt”) of the tick consisting of the smell of mammal sweat, fur-like textures and body heat, allowing it to detect a blood host, locate its skin and retrieve the nutrition needed to reproduce (Uexküll 2001). Furthermore, Uexküll argued for the mutual and historical constitution of parts in an organism as a prerequisite for the

¹¹ It should be noted that many concepts deployed in these theories have cybernetic roots and refer to *functionally* circular processes. This does not necessarily render them *definitionally* circular. A difference sometimes missed by philosophers biased towards logical hierarchies and linearity.

functional coupling between infrastructure and cognitive control (Uexküll 2001, Sørensen 2002a).

A more recent and famous example of systemic naturalist approaches is the theory of *autopoiesis* developed by the Chilean neurobiologists Humberto Maturana and Francisco Varela in the late 1960s (Maturana & Varela 1980, 1987; cf. Ziemke 2007).¹² Autopoiesis denotes the physical and cognitive self-constituting processes that make up a living system. The theory of autopoiesis takes the cognition-viability correlation to its extreme by not only stressing the dependence of cognition on the functional coupling arising in the structural self-organization of adaptive systems but by actually insisting on the *coextension* of the two aspects (Wheeler 1997, Boden 2000, Ziemke 2007). All living systems are consequently also cognitive according to the theory of autopoiesis. This claim is naturally highly controversial. Even if cognitive capacities are granted a very fundamental status in biology, few will grant them coextension with life (cf. Ziemke 2007). It seems less controversial to conceive of cognition and agency as coextensional, thus granting some level of autonomy to all life, but only cognition and agency to animal life forms.

For the remainder of this section, we will focus on another closely related contemporary theory of agency, the *interactivist* approach in which cognition and agency are seen as less universal but more hierarchically graded capacities than in the autopoietic model.¹³ According to the interactivist theory, agency denotes an emergent integrative and self-organizing capacity found solely in adaptive systems. Agency relates very closely to concepts of autonomy (Christensen & Hooker 2001) and self-maintenance (Bickhard & Terveen 1995).¹⁴ Complex adaptive systems are governed by self-organized norms emerging from the interactive dynamics of the system constituents and: “possess a process organization that, in interaction with the environment, performs work to guide energy into the processes of the system itself” (Christensen & Bickhard 2002). As such, interactivism is ontologically neutral and treats

¹² The theory of autopoiesis has strong similarities with Uexküll’s work (Ziemke & Sharkey 2001, Ziemke 2001).

¹³ Bickhard (unpublished) provides a brief introduction to the interactivist approach and Bickhard (in preparation) compares interactivism to the theory of autopoiesis.

¹⁴ There is a slight difference in the interactivist models presented. ‘Autonomy’ exclusively designates the special dynamic cohesiveness of adaptive systems whereas ‘self-maintenance’ includes non-recursively self-maintaining systems only contributing to their own self-maintenance under very restricted conditions. See below.

cells, organs, organisms, groups, and societies to be examples of complex adaptive systems at different levels.

What is special about cognitive adaptive systems, the ones to which agency is normally attributed, is the integrative nature of their self-governance (Christensen 2004). In fact, this greater integrative capacity together with pro-activity signifies autonomous systems of the highest complexity and is what agency amounts to. Again, integrative and pro-active capacities are graded qualities and do not allow for sharp definitional boundaries for agency.

According to a naturalist continuous perspective, cognition is explained as emergent from more fundamental biological self-maintaining capacities and as contributing to the self-maintenance of the particular system. Cognition provides increased integrative and pro-active capacities, primarily through the external ability to integrate tools and signs for self-maintaining purposes (Bickhard 1992, Dennett 1995, Sterelny 2004).¹⁵ Due to such “epistemic agency” (Sterelny 2004), cognitive systems gradually transcend dependence on the correlation between needs and the opportunities which evolution has provided. Among the more flexible ways of conducting self-maintenance belong planning, postponing certain kinds of need-fulfillment and sustaining goals. Contrary to classical rationalist conceptions of agency (e.g. Aristotle’s unmoved mover), agency does not amount to causal independence or originality of action but to organizational flexibility pushing the envelope of interactive possibilities.

3.2 Agency fuelled by norms

A fundamental notion in interactivist theory is the functional normativity by which adaptive systems are governed. Interactivism takes its departure from thermodynamically open systems to provide a naturalist explanation of how norms arise spontaneously in nature. Being far from equilibrium, such open systems depend on an on-going controlled input of materials, energy and information to keep up their functional organization and systemic cohesiveness. This dependence on input creates an asymmetry of normativity for the system; some things contribute to their self-maintenance and others do not.

¹⁵ It should be noted that the integrational perspective renders the internal-external distinction within epistemology somewhat arbitrary as features are measured by their contribution to the organization of dynamics and not spatial origin. Internal and external designate are – at best – graded qualities and not a sharp distinction.

Even though norms govern the interactions of the system, they are not all explicit, let alone conceptual, for the system itself. Many norms are implicit and not immediately identifiable by the system. Yet, it does make a difference, for example, for an organism whether it is hungry or not, whether the environment is hostile or not, etc. Sooner or later, it will have to take action to remove any 'deficit' in order to sustain life. The interactivist theory thus explains the basis of norms in causal systemic terms as variant constraints on the functional cohesiveness in open systems.

Norms both arise and reside endogenously in adaptive systems and are used for interactive guidance (input management) as interaction is cyclic and always for the sake of feedback – not the action *per se*. Adaptive systems have inherited traits which allow for simple spontaneous behaviors, but most behavior is guided by retaining the outcome of earlier interactions, successes or failures, which later function as normative anticipatory cues for subsequent actions. Some cues are infrastructurally 'wired' such as fixating on face-like forms in human infants (Johnson 1991). Others are conditioned response patterns such as the reflexive removal of a hand from something hot. Others again are abstract concepts exclusively found in human cognition.

Norms relate to the self-maintenance of the system, and anticipatory cues for possible interactive outcomes are constructed by the system itself. There is no such thing as external norms in this regard, since systems are unique as to which interactions support their self-maintenance. Energy in the form of food is of course a universal requirement for organisms but the kind and amount of food needed varies enormously. When it comes to information of relevance to the system in question, differences are even greater (cf. Uexküll's tick). Values nevertheless relate to the consequences of interaction for the system and hence on external feedback.

Adaptive systems have different dynamic means for self-maintenance which can roughly be distinguished as long-term or phylogenetic adaptation and short-term or ontogenetic and epigenetic adaptation. In this context, the manner in which organisms adapt epigenetically, i.e. by development and learning, is relevant. Research in developmental biology suggests that the majority of maturation is governed by both the genome and the actual environment (Oyama *et al.* 2001). As another example of increased awareness of processes, genes are no longer considered an exclusive and discrete source of developmental directives, but rather as one resource in an immensely

complex epigenetic development in interaction with an environment (Keller 2002, Depew & Weber 1997).

Through the ability to adjust responses on the fly, to meet changing needs and opportunities, adaptive (autonomous) systems differ from merely self-maintaining systems by being *recursively* self-maintaining. Recursive self-maintenance or adaptivity requires the capacity for error-detection and self-repair to change interactive strategies in the face of changed circumstances, in contrast to systems that only contribute to their self-maintenance in fixed circumstances. As an example of a minimally self-maintaining system, a burning candle provides fuel for the flame by melting wax, keeping an above-threshold combustion temperature for the flame to burn and attract oxygen and disposing waste by convectional air turbulence. But, it cannot create more wax when burned down or move if oxygen supplies run critically low. Candlelight is self-maintaining in a specific context but it is not autonomous in the same strong sense as agents that actively satisfy their inner norms for self-maintenance.

Both short- and long-term adaptation exploit the same ordering principles, namely variation and selection cycles. In fact, variation and selection are the fundamental ordering principle of self-organization (Bickhard & Campbell 2003, Campbell 1960). Even though this principle is best known as natural selection, it is also the active principle at other levels of organization in a range of complex systems. For instance, evidence in neuroscience suggests that variation and selection dynamics are fundamental in the brain as well. Multiple neuronal groups offer assistance in a given task and some get selected. The candidates chosen for a given task undergo synaptic strengthening (through neurological reinforcement) and, by training successful groups, become dominant (Edelman 1992).

4. Agency and AI

Agency has become a core notion in the fields of AI and robotics, especially since the rise of behavior-based robotics starting in the mid-1980s,¹⁶ which paralleled and further boosted a similar tendency towards *situated* and *embodied* cognition (e.g. Suchman 1987, Varela *et al.* 1991, Clancey 1997, Clark 1997, Sørensen 2002b, Ziemke 2002) in cognitive science more broadly. What is now called good old-fashioned AI (GOF AI)

¹⁶ The pioneering work of W. Grey Walter during the 1950s is often mentioned as the first example of this approach but it had little impact because of the dominance of knowledge-based AI (see below).

had been practicing a formalistic top-down approach to modeling intelligence for decades with great self-confidence but little success. By the late 1980s, roboticists and artificial life researchers such as Randall Beer, Luc Steels, Rolf Pfeifer, and, most prominently, Rodney Brooks redirected much AI research by shifting focus to how relatively simple organisms manage to negotiate their surroundings for the purpose of self-maintenance (e.g. Beer 1990, Steels 1994, Pfeifer & Scheier 1999, Brooks 1999). GOFAI's intellectualistic understanding of agency as a fully transparent, volitional and rational self-control gave way for a view of agency as an opportunistically *ad hoc*, but global (autonomous) capacity. Intelligent agency, thus understood, indicates the range of environmental conditions a system is able to manage without external support and is sometimes called a horizontal notion of intelligence. In contrast, 'vertical' kinds of intelligence such as expert systems and the superior chess master program *Deep Blue*, which had primarily interested GOFAI, became construed as less intelligent than even the simplest organism with neither autonomy nor agency (e.g. Pfeifer & Scheier 1999).

4.1 Autonomous agents

Among the 'New AI' researchers, autonomy and agency have come to be two of the defining terms of the field and some effort has been devoted to conceptual discussions of these fundamental terms. Nevertheless they are often used in a vague or intuitive sense that emphasizes certain surface similarities between living and robotic systems, and takes these as grounds for transferring properties such as autonomy and agency from the former to the latter class of systems, without actually defining what exactly these properties are (cf. Ziemke 2007). Beer (1995), for example characterized 'autonomous agents' as follows:

By *autonomous agent*, we mean any embodied system designed to satisfy internal or external goals by its own actions while in continuous long-term interaction with the environment in which it is situated. The class of autonomous agents is thus a fairly broad one, encompassing at the very least all animals and autonomous robots. (Beer 1995: 173)

As the present widespread use of terms such as 'software agents' and 'internet agents' indicates, the defining characteristics of (physical) embodiment or basic autonomy are downplayed in some fields (cf. Chrisley & Ziemke 2002, Ziemke 2003). In fact, most practitioners have fallen back into a more formal control-theoretical way

of interpreting autonomy and agency as the capacity of negotiating a given environment (virtual or real) on the basis of system-internal rules without requirements for the genesis of such internal rules (cf. Ziemke 1998).¹⁷ In a paper specifically devoted to scrutinizing the multifarious meanings, the concept ‘agent’ had acquired by the mid-1990s, Stan Franklin and Art Graesser provided the following definition of autonomous agents:

An autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future. (Franklin & Graesser 1996)

In Franklin and Graesser’s definition, ‘situated’ equals placement (like apples in a basket), ‘environment’ means context in a widest sense, ‘sensing’ amounts to mere registering and ‘own agenda’ does not require that the goals actually in any way concern the system itself. They note that the definition include thermostats, but argue that this is due to the fact that their definition only captures the essence of agency. Even if Franklin and Graesser’s liberal notion of autonomous agency is controversial among AI researchers, it demonstrates the plasticity of the term ‘agent’ due to the lack of a common understanding. In fact, it is hard to see that deploying ‘autonomous agent’ as suggested serves any purpose, but legitimizing its application to fields such as software agents.

However, other sub-fields of AI research, especially within robotics, taking the new ideas more seriously have slowly emerged. Agreeing on the importance of how autonomous systems *acquire* and dynamically maintain internal control mechanisms, new fields such as evolutionary and epigenetic robotics have emerged (e.g. Nolfi & Floreano 2000, Berthouze & Ziemke 2003). By focusing on growth, development and learning these new and explicitly biologically inspired fields investigate the self-organization of structural-functional integration in adaptive systems, putting as much emphasis on the infrastructure and morphology of ‘autonomous’ systems as their actual control mechanisms (software).

As pointed out above, the attribution of agency to artificial systems, such as robots, artificial life forms or software agents, hinges much on certain surface

¹⁷ This tendency is not only due to ignorance but a new pragmatic stance on artificial intelligence. Whereas some parts of the AI turned to biological models to create ‘strong AI’ after the failure of GOFAI, others have re-named their approach ‘computational intelligence’ to indicate a focus primarily on practically feasible and commercially valuable semi-autonomous IT without philosophical considerations.

similarities, i.e. properties that are considered characteristic for living systems which are taken to be natural autonomous agents and can be attributed, by analogy, to their presumed artificial counterparts. Some of these properties are: some form of autonomy, some form of 'life', 'situatedness' (essentially interacting with an environment), 'embodiment' (although it is increasingly unclear what that means, cf. Chrisley & Ziemke 2002, Ziemke 2003), goal-directed behavior, self-steering, self-maintenance (in some sense), and capacities for adaptation, development and learning (cf. Sharkey & Ziemke 1998). It is fairly obvious that few, if any, of these are exactly the same in living and artificial 'agents'. For example, few would deny that there are a number of significant differences between living bodies and robot bodies (cf. Ziemke 2007).

That leaves us with three distinct possibilities. Firstly, it might be argued that in fact there are no such things as 'agents'. As pointed out above, much science has in some sense been about eliminating agency from scientific explanations and reducing it to lower-level processes, such that agency might after all turn out to fall into the same category as the phlogiston, the *élan vitale*, and perhaps the soul. If so, then agency might still be useful to attribute to both living and artificial systems alike in folk-psychological explanations, but such explanations cannot be considered to have any significant scientific value. Secondly, it might be argued that the above properties, or perhaps a subset of them, are in fact all that agency amounts to. If so, then animals, if not all living systems, and robots are all autonomous agents in roughly the same sense, as Beer's as well as Franklin and Graesser's above definitions of autonomous agents imply, and as many AI researchers seem to assume. The third possibility, however, which we argue for in this paper, is that the differences between living (animal) and (current) non-living systems are actually crucial to agency.

5. Implications for artificial agency

Considering the notion of agency put forward in section 3, it seems clear that no existing artificial system presently fulfills these strict requirements. There have been no examples of physical growth or self-maintenance in the strong sense suggested in robots and artificial life forms so far (cf. Sørensen 2002a, Ziemke & Sharkey 2001, Ziemke 2001). In fact, it might seem virtually impossible to ever create artificial agency given the dynamic and infrastructural requirements listed above. So, have we simply

replaced an anthropocentric rationalism with bio-chauvinism by insisting so much on the structural integration that is characteristic of living organisms?

Even if approaches related to the interactivist model are sometimes criticized for being overly organism-focused, the focus is in fact on specific self-maintaining processes and not privileged substances. Organisms merely serve as paradigmatic examples and the interactivist notions of agency and autonomy are not by definition restricted to biological organisms.¹⁸ Christensen & Hooker (2000) and Christensen & Bickhard (2002), for example, mention species and colonies as biological examples of other types of autonomous systems and societies and economies as non-biological examples. Similarly, Maturana and Varela pointed out that for autonomous systems “the phenomena they generate in functioning as autopoietic unities depend on their organisation and the way this organisation comes about, and not on the physical nature of their components” (Maturana & Varela 1987). Instead, the defining characteristic of systems governed by agency is their specific organization of system-maintaining processes.

Hence, there is nothing in the presented modern theories of adaptive systems that rules out the possibility of artificial autonomy and agency. Yet, as discussed in more detail elsewhere (Ziemke 2001), a major problem with current ‘New AI’ and adaptive robotics research is that, despite its strong biological inspiration it has focused on establishing itself as a new paradigm *within* AI and cognitive science, i.e. as an alternative to the traditional computationalist paradigm. Relatively little effort has been made to make the connection to other theories addressing issues of autonomy, situatedness and embodiment, although not necessarily under those names. More specifically, New AI distinguishes itself from its traditional counterpart in its interactive view of *knowledge*. In particular, recent work in the field of adaptive robotics, as discussed above, is largely compatible with the interactivist or radically constructivist view (e.g. von Glasersfeld 1995) of the construction of knowledge through sensorimotor interaction with the environment with the goal of achieving some ‘fit’ or ‘equilibrium’ between internal behavioral/conceptual structures and experiences of the environment. However, the organic roots of these processes, which were emphasized

¹⁸ To be fair however, the step from paradigmatic examples over general descriptive bias to default ontology is always lurking in scientific enthusiasm and popularity. Just think of the computer metaphor for the brain, which went from being a didactic example to being a doctrine dominating cognitive science for many years.

in the theoretical biology of von Uexküll or Maturana and Varela's theory of autopoiesis, are often ignored in New AI, which still operates with a view of the body that is largely compatible with mechanistic theories and a view of control mechanisms that is still largely compatible with computationalism. This means that the robot body is typically viewed as some kind of input- and output-device that provides physical grounding to the internal computational mechanisms (cf. Ziemke 2001).

Thus, in practice, New AI has become a theoretical hybrid, or in fact a 'tribrid', combining a mechanistic view of the body with the interactivist/constructivist notion of interactive knowledge, and the functionalist/computationalist hardware-software distinction and its view of the activity of the nervous system as computational (cf. Ziemke 2001). The theoretical framework of interactivism, as elaborated above, might serve as a useful starting point for a conceptual defragmentation of current 'embodied' AI research in general and for further progress towards artificial agency in particular.

On a more positive note, a renewed pursuit of AI incorporating structural aspects such as dynamic materials as prerequisites for cognitive systems could bring embodied cognitive science beyond lip service and contribute to a necessary opposition to millennia of dualist hegemony. AI could very well become an important pioneering field for a unified approach to cognition if systematically investigating self-organizing and –maintaining capacities in reconfigurable and 'growing' new synthetic materials (Sørensen 2004, 2005, in preparation).

6. Concluding remarks

Historically, the blurry concept of agency has played a pivotal role in theories of man and his world. Despite a general scientific abandonment of agent-based and substantialist theories, the notion of the agent is still very dominant in most humanistic theories and in our self-understanding. And even if not completely without ontological merit and scientific value, the notion of agency employed in many theories still needs thorough revision.

In an effort to clarify and qualify the concept of agency, we have examined different notions of agency ranging from a heavily 'entitled' folk conception to the radically distributed and process-emergent models in dynamic systems theories. Historically, the notion of agency has been wed to autonomous rationality as a self-contained and strictly human (and divine) capacity as exemplified by Aristotle's

unmoved mover. In AI, the concept of agency is mostly defined by equally vague and ill-understood concepts of autonomy and embodiment and mostly opportunistically to fit a specific engineering goal. In modern naturalistic cognitive theories on the other hand – particularly in interactivist and autopoietic theories – ‘agency’ tends to denote the capacity to orchestrate the self-maintaining flow of energy; a capacity identical in principle but not in complexity for humans and other animals. Agency is understood as an integrational organization *capacity* of open systems and not as a uniform *principle*. In fact, rather than being a primitive *explanans*, agency is a highly complex *explanandum*. Hence, acknowledging the variety of levels and types of agency, the use of the concept should be more carefully considered than hitherto by explicating the specific meaning intended.

In relation to the field of AI we have argued, despite much conceptual progress in AI research towards acknowledging the systemic foundations of intelligence, agency is mostly used in an almost vacuously broad sense carrying only superficial similarities with known natural examples. Most notions of agency still rest on Cartesian leanings toward a dual understanding of agents as divided into control mechanisms and structure. In addition to the conceptual obstacle brought on by a – basically dualistic – pre-occupation with software, the enterprise of creating artificial agency suffers from the lack of dynamic structures capable of integrating systemically with information processing capacities. The next breakthrough in the pursuit of true artificial intelligence is likely to come, at least partly, from new structural principles and so-called ‘smart materials’ capable of growth, development and self-organization.

So, even if human-level cognition is no longer the prime goal for AI research, the enterprise has not become any easier. We are in need of a much more fundamental understanding of how intelligence and agency arise in concert with integrated self-organizing capacities in adaptive systems. To obtain intelligence and agency by synthetic means, we must find ways to obey basic principles of self-maintenance. In the words of Rick Belew (1991): “The dumbest smart thing you can do is to stay alive”. Thus, in principle there are no obstacles for creating artificial agency but probably an immensely bumpy empirical road ahead.

On the other hand, the broad spectrum of AI-related research might nevertheless turn out to be crucially instrumental to the modeling and understanding of cognition. Given that cognition is a quite complex and elusively processual

phenomenon our understanding of it is likely to benefit significantly from extensive synthetic and empirical investigations. Taking the emergent nature of these matters serious there is a lot of truth to Braitenberg's "law of uphill analysis and downhill invention" (Braitenberg 1984). Acknowledging that the processes underlying complex phenomena are themselves mostly less complex and often quite different we should not throw out the AI baby with the fuzzy conceptual bathwater. AI research is definitely an important part of the cognitive sciences and it ought not to be relegated as 'mere engineering'. Yet, AI is also a field extraordinarily obliged to proceed carefully due to the great cultural interest it is enjoying, as exemplified by the prominence of AI in the science fiction genre.

Finally, it should be noted that biological models are currently impacting several scientific fields as well as culture in general (Sørensen 2003a, 2003b, 2004, 2005, in preparation). The theories and arguments put forward in this paper undoubtedly carry the mark of this general tendency. But even if the biological paradigm fades out when new scientific trends emerge, and life turns out to be an arbitrary level of reference for AI (cf. the graded notions of autonomy, agency etc.), there are, for the time being, strong inductive reasons to couple agency with the kind of integrated/integrative adaptive self-maintenance so far solely found in living systems.

Acknowledgements

Tom Ziemke is supported by a European Commission grant to the project "*Integrating Cognition, Emotion and Autonomy*" (ICEA, IST-027819, www.his.se/icea) as part of the European *Cognitive Systems* initiative.

References:

- Auyang, S. Y. (1998). 'Foundations of Complex-system Theories'; in *Economics, Evolutionary Biology, and Statistical Physics*. Cambridge: Cambridge University Press.
- Beer, R. D. (1990). *Intelligence as Adaptive Behavior: An experiment in computational neuroethology*. Boston: Academic Press.
- Beer, R. D. (1995). 'A dynamical systems perspective on autonomous agents'. *Artificial Intelligence*, 72: 173-215.
- Belew, R.K. (1991). 'Artificial Life: a constructive lower bound for Artificial Intelligence'. *IEEE Expert* 6(1): 8-14, 53-59.
- Berthouze, L. & Ziemke, T. (eds.) (2003). 'Epigenetic Robotics – Modelling Cognitive Development in Robotic Systems' (special issue). *Connection Science*, 15(4).
- Bickhard, M. H. & Terveen, L. (1995). *Foundational Issues in Artificial Intelligence and Cognitive Science - Impasse and Solution*. Amsterdam: Elsevier Scientific.

- Bickhard, M. H. (unpublished). 'Interactivism. A manifesto'.
<http://www.lehigh.edu/~mbb0/pubspage.html>
- Bickhard, M. H. (1992). 'Scaffolding and Self-Scaffolding: Central Aspects of Development'; in L. T. Winegar & J. Valsiner (eds.), *Children's Development within Social Contexts: Research and Methodology*, Vol 2: 33-52. Hillsdale: Erlbaum.
- Bickhard, M. H. (in preparation). *The Whole Person. Toward a Naturalism of Persons —Contributions to an Ontological Psychology*.
- Bickhard, M. H. & Campbell, D. T. (2003). 'Variations in Variation and Selection: The Ubiquity of the Variation-and-Selective Retention Ratchet in Emergent Organizational Complexity'. *Foundations of Science*, 8 (3): 215-282.
- Boden, M.A. (2000). 'Autopoiesis and life'. *Cognitive Science Quarterly* 1: 117—145.
- Bonabeau et al. (1998). *Swarm Intelligence, From natural to Artificial Systems* (SFI Studies in the Sciences of Complexity). Oxford: Oxford University Press.
- Braitenberg, V. (1984). *Vehicles, experiments in synthetic psychology*. Cambridge: MIT Press.
- Brooks, R.A. (1999). *Cambrian Intelligence: The early history of the new AI*. Cambridge: MIT Press.
- Browning, D. & Myers, W. T. (1998). *Philosophers of Process*. New York: Fordham University Press.
- Campbell, D. T. (1973). 'Ostensive Instances and Entitativity in Language Learning'; in W. Gray & N. D. Rizzo. (eds.), *Unity Through Diversity*, vol. 2. New York: Gordon and Breach.
- Campbell, D. T. (1960). 'Common Fate, Similarity, and Other Indices of the Status of Aggregates of Persons as Social Entities'. *Behavioral Sciences* 3: 14-25.
- Campbell, R. J. & Bickhard, M. H. (2000). 'Physicalism, Emergence, and Downward Causation'; in P. Andersen et al. (eds.), *Downward Causation. Minds, Bodies and Matter*. Århus: Aarhus University Press.
- Clancey, W. J. (1997). *Situated Cognition: On Human Knowledge and Computer Representations*. New York: Cambridge University Press.
- Clark, A. (1997). *Being There: Putting Brain, Body and World Together Again*. Cambridge: MIT Press.
- Clark, A. (2003). *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Cambridge: Oxford University Press.
- Chrisley, R. & Ziemke, T. (2002), 'Embodiment'; in *Encyclopedia of Cognitive Science*, pp. 1102-1108. London: Macmillan.
- Christensen, W. D. (2004). 'Self-directedness, integration and higher cognition'. *Language Sciences* 26: 661-692.
- Christensen, W. D. & Bickhard, M. (2002). 'The Process Dynamics of Normative Function'. *Monist*, vol. 85, no. 1: 3-28.
- Christensen, W. D. & Hooker, C. A. (2001). 'Self-directed agents'; in J. McIntosh (ed.), *Naturalism, Evolution, and Intentionality*, *Canadian Journal of Philosophy, Special Supplementary Volume* (27).
- Christensen, W.D. & Hooker, C. A. (2000). 'Anticipation in autonomous systems: foundations for a theory of embodied agents'. *International Journal of Computing Anticipatory Systems*, Volume 5: 135-154.
- Dennett, D. C. (1995). *Darwin's Dangerous Idea*. New York: Simon and Schuster.
- Depew, D. J. & Weber, B. H. (1997). *Darwinism Evolving: Systems Dynamics and the Genealogy of Natural Selection*. Cambridge: MIT Press.
- Edelman, G. M. (1992). *Bright Air, Brilliant Fire: In the Matter of the Mind*. New York: Basic Books.
- Franklin, S. & Graesser, A. (1997). 'Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents'. *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages*. Heidelberg: Springer-Verlag.
- Gallagher, S. (2000). 'Philosophical conceptions of the self: implications for cognitive science'. *Trends in Cognitive Sciences*, 4(1): 14-21.
- Godfrey Smith, P. (1998). *Complexity and the Function of Mind in Nature*. Cambridge: Cambridge University Press.
- Goldberg, E. (2002). *The Executive Brain: Frontal Lobes and the Civilized Mind*. Cambridge: Oxford University Press.
- Hendriks-Jansen, H. (1996). *Catching Ourselves in the Act: Situated Activity, Interactive Emergence, Evolution, and Human Thought*. Cambridge: MIT Press.

- Hutchins, E. (1995). *Cognition in the Wild*. Cambridge: MIT Press.
- Johnson, M. H. Morton, J. (1991). *Biology and cognitive development: The case of face recognition*. Oxford: Basil Blackwell.
- Keller, E. F. (2002). *The Century of the Gene*. Boston: Harvard University Press.
- Nolfi, S. & Floreano, D. (2000). *Evolutionary Robotics*. Cambridge: MIT Press.
- Oyama, S. et al. (eds.) (2001). *Cycles of Contingencies: Developmental Systems and Evolution*. Cambridge: MIT Press.
- Petitot, J. et al. (eds.) (2000). *Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science*. Stanford: Stanford University Press.
- Pfeifer, R. & Scheier, C. (1999). *Understanding Intelligence*. Cambridge: MIT Press.
- Russell, S. J. & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, NJ: Prentice Hall.
- Sharkey, N. E. & Ziemke, T. (1998). 'A consideration of the biological and psychological foundations of autonomous robotics'. *Connection Science*, 10(3- 4): 361-391.
- Sharkey, N. E. & Ziemke, T. (2001). 'Mechanistic vs. Phenomenal Embodiment: Can Robot Embodiment Lead to Strong AI?' *Cognitive Systems Research*, 2 (4): 251-262.
- Smolin, L. (2003). 'Loop Quantum Gravity'; in J. Brockman (ed.), *The New Humanist: Science at The Edge*. New York: Barnes & Noble.
- Steels, L. (1994). 'The Artificial Life Roots of Artificial Intelligence'. *Artificial Life*, 1: 75-100.
- Sterelny, K. (2001). 'Niche Construction, Developmental Systems, and the Extended Replicator'; in S. Oyama et al. (eds.), *Cycles of Contingencies. Developmental Systems and Evolution*. Cambridge: MIT Press.
- Sterelny, K. (2004). 'Externalism, Epistemic Artefacts and The Extended Mind'; in: R. Schantz (ed), *The Externalist Challenge. New Studies on Cognition and Intentionality*. New York: Mouton de Gruyter.
- Suchman, L. A. (1987). *Plans and Situated Action: The Problem of Human-Machine Communication*. New York: Cambridge University Press.
- Susi, T. & Ziemke, T. (2001). 'Social Cognition, Artifacts, and Stigmergy'. *Cognitive Systems Research*, 2(4): 273-290.
- Sørensen, M. H. (2003a). 'Assistive Ecologies. Biomimetic Design of Ambient Intelligence'. *Proceedings, Intelligent Agent Technologies*, Halifax, 2003.
- Sørensen, M. H. (2005). *Ambient Intelligence Ecologies. Toward Biomimetic IT*. Ph.D. Dissertation, IT University of Copenhagen.
- Sørensen, M. H. (in preparation., 'Design Symbiosis: Dynamic Design of IT').
- Sørensen, M. H. (2002a). 'Fra mider til androider'. *Semikolon* 3
- Sørensen, M. H. (2003b). 'It's A Jungle Out There: Toward Design Heuristics for Ambient Intelligence Ecologies'. *Proceedings, Computer, Communication and Control Technologies*, Orlando, 2003.
- Sørensen, M. H. (2002b). 'The Body is Back'. *Connection Science Journal*, Vol. 14, nr. 1
- Sørensen, M. H. (2004). 'The Genealogy of Biomimetics: Half a Century's Quest for Dynamic IT'. *Proceedings, The First International Workshop of Biologically Inspired Approaches to Advanced Information Technology*, Lausanne, 2004.
- Uexküll, J. v. (2001). 'An introduction to Umwelt'; in K. Kull (ed.): *Semiotica – Special Issue: Jakob von Uexküll: A paradigm for biology and semiotics*. Haag: Mouton de Gruyter.
- Varela, F.J. (1979). *Principles of Biological Autonomy*. New York: Elsevier.
- Varela, F.J. et al. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge: MIT Press.
- Varela, F.J. (1997). Patterns of Life: Intertwining Identity and Cognition. *Brain and Cognition*, 34: 72-87.
- Wheeler, M. (1997). 'Cognition's Coming Home: the Reunion of Life and Mind'; in P. Husbands & I. Harvey (eds.). *Proceedings of the Fourth European Conference on Artificial Life*. 10-19. Cambridge: MIT Press.
- von Glasersfeld, E. (1995). *Radical Constructivism – A Way of Knowing and Learning*. London: Falmer Press..

- Wimsatt, W. C. (forthcoming), *Re-engineering Philosophy for Limited Beings: Piecewise Approximations To Reality*. Cambridge: Harvard University Press
- Ziemke, T. (2001). 'The Construction of 'Reality' in the Robot: Constructivist Perspectives on Situated AI and Adaptive Robotics'. *Foundations of Science*, 6(1): 163-233.
- Ziemke, T. (2003). 'What's that thing called embodiment?'; in *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*. Mahwah: Lawrence Erlbaum.
- Ziemke, T. (2007) 'What's life got to do with it?'. in A. Chella & R. Manzotti, (eds.), *Artificial Consciousness*. Exeter, UK: Imprint Academic.
- Ziemke, T. (ed.) (2002). 'Situated and Embodied Cognition' (special issue). *Cognitive Systems Research*, 3(3).
- Ziemke, T. & Sharkey, N. (2001). 'A stroll through the worlds of robots and animals: Applying Jakob von Uexküll's theory of meaning to adaptive robots and artificial life'. *Semiotica*, 134(1-4): 701-746.

Paper received April 2004; revised February 2007

SUBSCRIBE

COGNITIVE SEMIOTICS

Editorial board:

Line Brandt (Aarhus), Per Aage Brandt (Cleveland), Frank Kjørup (Copenhagen), Todd Oakley (Cleveland), Jacob Orquin (Aarhus), Jakob Simonsen (Aarhus), and Jes Vang (Aarhus).

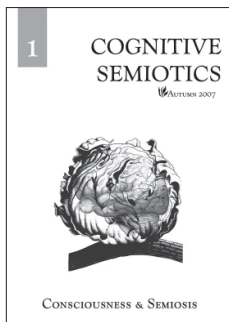
The first of its kind, *Cognitive Semiotics* is a multidisciplinary journal devoted to high quality research, integrating methods and theories developed in the disciplines of cognitive science with methods and theories developed in semiotics and the humanities, with the ultimate aim of providing new insights into the realm of human signification and its manifestation in cultural practices. Accordingly, readers will have the opportunity to engage with ideas from the European and American traditions of cognitive science and semiotics, and to follow developments in the study of meaning – both in a cognitive and in a semiotic sense – as they unfold internationally.

The initiative to create a transatlantically based journal comes from the Center for Cognition and Culture at the department of Cognitive Science at Case Western Reserve University (Cleveland), and from a group of researchers, based in Aarhus and Copenhagen, trained in cognitive semiotics at the Center for Semiotics at the University of Aarhus, and in language and literature at the University of Copenhagen. By bringing together scholars from multiple disciplines, the editors hope to provide a revitalized perspective on the semiotic field.

The printed journal will be accompanied by a comprehensive online resource, featuring novel content, newsletters and discussion forums: www.cognitivesemiotics.com.

Cognitive Semiotics is published twice a year, in April and October.

ISSN 1662-1425



Issue 1 • Autumn 2007

CONSCIOUSNESS & SEMIOSIS

This issue will contain contributions by:

Liliana Albertazzi	Svend Østergaard
Bernard Baars	Jean-Luc Petit
Per Aage Brandt	Ernst Pöppel
Terrence Deacon	Frederik Stjernfelt
Anthony Jack	Patrizia Violi

You will find an order form on the reverse side of this page.

Please send your order to:

Peter Lang AG · International Academic Publishers
Moosstrasse 1 · P.O. Box 350 · CH-2542 Pieterlen · Switzerland
Tel.: +41 32 376 17 17 · Fax: +41 32 376 17 27 · info@peterlang.com · www.peterlang.com

SUBSCRIBE

ORDER FORM

I would like to subscribe to **COGNITIVE SEMIOTICS**

Subscription price (2 issues):

sFr. 61.00 / €* 43.00 / €** 44.00 / € 40.00 / £ 26.00 / US-\$ 49.00

I would like to order a single issue of **COGNITIVE SEMIOTICS**

Issue 1 · Autumn 2007 (Art. No. 81602):

sFr. 42.00 / €* 28.90 / €** 29.70 / € 27.00 / £ 18.00 / US-\$ 34.00

Invoice Eurocard/MasterCard VISA

Credit card number

CVV/CVC

/

Exp. date

Signature

Name

Address

Date

Signature

Prices are subject to change without notice and do not include shipping and handling.
* includes VAT - valid for Germany ** includes VAT - valid for Austria

Please send your order to:

Peter Lang AG · International Academic Publishers
Moosstrasse 1 · P.O. Box 350 · CH-2542 Pieterlen · Switzerland
Tel.: +41 32 376 17 17 · Fax: +41 32 376 17 27
info@peterlang.com · www.peterlang.com

European Semiotics: *Language, Cognition, and Culture* Sémiotiques Européennes : *langage, cognition et culture*

Series edited by / Collection rédigée par
Per Aage Brandt (Aarhus), Wolfgang Wildgen (Bremen/Brême),
and/et Barend van Heusden (Groningen/Groningue)

European Semiotics originated from an initiative launched by a group of researchers in Semiotics from Denmark, Germany, Spain, France and Italy and was inspired by innovative impulses given by René Thom and his "semiophysics". The goal of the series is to provide a broad European forum for those interested in semiotic research focusing on *semiotic dynamics* and combining *cultural, linguistic and cognitive perspectives*.

This approach, which has its origins in Phenomenology, Gestalt Theory, Philosophy of Culture and Structuralism, views semiosis primarily as a cognitive process, which underlies and structures human culture. Semiotics is therefore considered to be the discipline suited *par excellence* to bridge the gap between the realms of the Cognitive Sciences and the Sciences of Culture.

The series publishes monographs, collected papers and conference proceedings of a high scholarly standard. Languages of publication are mainly English and French.

Sémiotiques européennes est le résultat d'une initiative prise par un groupe de chercheurs en sémiotique, originaires du Danemark, d'Allemagne, d'Espagne, de France et d'Italie, inspirée par l'impulsion innovatrice apportée par René Thom et sa « sémiophysique ». Le but de cette collection est de fournir une tribune européenne large à tous ceux qui s'intéressent à la recherche sémiotique portant sur *les dynamiques sémiotiques*, et réunissant des *perspectives culturelles, linguistiques et cognitives*.

Cette approche, qui combine différentes sources, telle que la phénoménologie, le gestaltisme, la philosophie de la culture et le structuralisme, part du principe que la sémiosis est essentiellement un procès cognitif, qui sous-tend et structure toute culture humaine. Dans cette approche, la sémiotique est donc considérée comme la discipline par excellence capable de créer un pont entre les domaines des sciences cognitives et ceux des sciences de la culture.

Sémiotiques européennes accueille tant des monographies que des anthologies et des actes de colloques d'un haut niveau de recherche, rédigés de préférence en anglais et en français.

Volume 1: Wolfgang Wildgen

De la grammaire au discours

Une approche morphodynamique
Préface de René Thom

Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Wien, 1999. XI, 351 p.

ISBN 978-3-906762-03-6 br.

sFr. 77.- / €* 56.80 / €** 58.40 / € 53.10 / £ 34.- / US-\$ 63.95

Le paradigme morphodynamique a été introduit par René Thom vers 1970 et fut élaboré par l'auteur pendant deux décennies. Ce livre présente le paradigme et propose des applications pour tous les domaines majeurs de la linguistique contemporaine. Les chapitres centraux traitent des problèmes du lexique, de la morphologie, de la syntaxe, de l'analyse textuelle et de l'énonciation. Un appendice donne une introduction à la mathématique qualitative. Cet ouvrage répond à une demande légitime : un paradigme nouveau doit montrer son efficacité et son originalité dans tous les secteurs d'une discipline afin de remplacer les paradigmes désuets.

Volume 2: Lene Fogsgaard

Esquemas copulativos de SER y ESTAR

Ensayo de semiolingüística

Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Wien, 2000. 382 pp.

ISBN 978-3-906764-22-1 en rústica

sFr. 89.- / €* 65.70 / €** 67.50 / € 61.40 / £ 40.- / US-\$ 73.95

El presente libro constituye un estudio de las estructuras copulativas en español mediadas por los verbos *ser* y *estar* como expresión de dos modos ónticos. El enfoque es semiolingüístico y pone en funcionamiento un modelo dinámico en espiral que hace posible la ubicación sistemática de las diferentes construcciones de *ser* y *estar*, sean éstas predicativas, copulativas o perifrásticas, presentándolas como discontinuidades de un continuo. El resultado de la investigación demuestra la existencia de un esquema cognitivo y una estructura enunciativa específica para sendos verbos. Esto permite ofrecer una explicación argumentada de la distribución de *ser* y *estar* facilitando así su aprendizaje.

Volume 3: Jean Petitot

Morphogenesis of Meaning

Translated by Franson Manjali

Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Oxford, Wien, 2004. XI, 279 pp., num. ill.

ISBN 978-3-03910-104-7 pb.

sFr. 70.- / €* 48.30 / €** 49.60 / € 45.10 / £ 32.- / US-\$ 53.95

The original French version of «Morphogenesis of Meaning» was one of the important breakthroughs in cognitive semiolinguistics during the eighties. It provided a deep philosophical elaboration of René Thom's Catastrophe Theory and exerted great influence on the main semiolinguistic schools in the world. Translated by Franson Manjali and revised with his help, the present English version focuses on dynamical modeling perspectives, the mathematical content of the models, and epistemologically foundational issues. The central problem dealt with is that of structure. Topological and dynamical models of morphogenesis are applied to structuralist theories such as phonology and categorical perception (Roman Jakobson), structural syntax and geometrical interpretation of case grammars (Lucien Tesnière, Louis Hjelmslev), narrative structures (Vladimir Propp, Claude Lévi-Strauss, Algirdas J. Greimas) as well as semiogenesis. It can be seen that there exists a theoretical convergence between many trends in American cognitive linguistics (Charles Fillmore, Len Talmy, Ron Langacker, George Lakoff) and the dynamical modeling perspectives which were developed in the context of European structuralism.

Volume 4: Per Aage Brandt

Spaces, Domains, and Meaning

Essays in Cognitive Semiotics

Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Oxford, Wien, 2004. 271 pp., num. fig.

ISBN 978-3-03910-227-3 pb.

sFr. 73.- / €* 50.30 / €** 51.70 / € 47.- / £ 33.- / US-\$ 55.95

Cognitive Semiotics is a new discipline dedicated to the analysis of meaning. It combines cognitive linguistics and semantics with structural and dynamic semiotics, and seeks to elaborate a coherent framework for the study of language and thought, gesture and culture, discourse and text, art and symbolization in general. The essays of this book develop a semiotic elaboration of the theory of mental spaces, a grounding hypothesis of semantic domains, and the methodologically necessary idea of a mental architecture corresponding to the neural organization of our brain, and compatible with the basic facts of human phenomenology. This volume presents the author's recent research, carried out at the Aarhus center, where American and European approaches to language-based semantics have been meeting for a particularly inspiring decade.

Volume 5: Marcel Bax / Barend van Heusden / Wolfgang Wildgen (eds.)

Semiotic Evolution and the Dynamics of Culture

Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Oxford, Wien, 2004. XIX, 318 pp., num. fig. and tables
ISBN 978-3-03910-394-2 pb.

sFr. 87.- / €* 60.- / €** 61.70 / € 56.10 / £ 40.- / US-\$ 66.95

This book is about patterns of development in the history of culture. Bringing together three areas of research: *semiotics*, *cultural history*, and *evolutionary psychology*, it attempts to bridge the gap that still separates the study of culture from the cognitive sciences. The multidisciplinary approach chosen by the contributors derives its impetus from the deep conviction that in order to understand the logic of cultural development, one must take the building blocks of culture, that is, signs and language, as a starting point for research. Central issues related to patterns of cultural evolution are dealt with in contributions on the development of mind and culture, the history of the media, the diversity of sign systems, culture and code, and the dynamics of semiosis. Theoretically oriented contributions alternate with in-depth case studies on such diverging topics as the evolution of language and art in pre-history, ritual as the fountainhead of indirect communication, developments in renaissance painting, the evolution of classification systems in chemistry, changing attitudes toward animal consciousness, and developments in computer technology.

Volume 6: Ángel López-García

The Grammar of Genes

How the Genetic Code Resembles the Linguistic Code

Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Oxford, Wien, 2005. 182 pp.

ISBN 978-3-03910-654-7 pb.

sFr. 55.- / €* 38.- / €** 39.10 / € 35.50 / £ 24.80 / US-\$ 42.95

Mankind is the only speaking species on earth. Hence language is supposed to have a genetic basis, no matter whether it relies on general intelligence, or on a linguistic module. This study proposes that universal formal properties of the linguistic code emerged from the genetic code through duplication. The proportion of segmental duplication is clearly higher in the human genome than in any other species, and duplication took place 6 million years ago when humans separated from the other hominid branches. The evolution of language is therefore supposed to be a gradual process with a break. This book describes a lot of striking formal resemblances the genetic code and the linguistic code hold in common. The book aims to reconcile generative grammar with cognitive semiotics showing that both of them constitute instances of embodiment.

Volume 7: Ole Kühn

Musical Semantics

Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Oxford, Wien, 2007. 261 pp.

ISBN 978-3-03911-282-1 pb.

sFr. 69.- / €* 47.60 / €** 49.- / € 44.50 / £ 28.90 / US-\$ 57.95

Music offers a new insight into human cognition. The musical play with sounds in time, in which we share feelings, gestures and narratives, has fascinated people from all times and cultures. The author studies this semiotic behavior in the light of research from a number of sources. Being an analytical study, the volume combines evidence from neurobiology, developmental psychology and cognitive science. It aims to bridge the gap between music as an empirical object in the world and music as lived experience. This is the semantic aspect of music: how can something like an auditory stream of structured sound evoke such a strong reaction in the listener? The book is in two parts. In the first part, the biological foundations of music and their cognitive manifestations are considered in order to establish a groundwork for speaking of music in generic, cross-cultural terms. The second part develops the semantic aspect of music as an embodied, emotively grounded and cognitively structured expression of human experience.

Ole Kühl

Musical Semantics

Music offers a new insight into human cognition. The musical play with sounds in time, in which we share feelings, gestures and narratives, has fascinated people from all times and cultures.

The author studies this semiotic behavior in the light of research from a number of sources. Being an analytical study, the volume combines evidence from neurobiology, developmental psychology and cognitive science. It aims to bridge the gap between music as an empirical object in the world and music as lived experience. This is the semantic aspect of music: how can something like an auditory stream of structured sound evoke such a strong reaction in the listener?

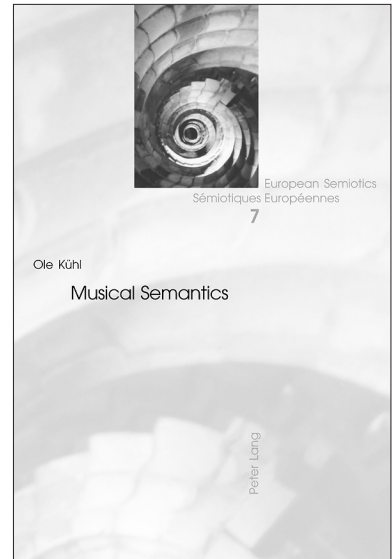
The book is in two parts. In the first part, the biological foundations of music and their cognitive manifestations are considered in order to establish a groundwork for speaking of music in generic, cross-cultural terms. The second part develops the semantic aspect of music as an embodied, emotively grounded and cognitively structured expression of human experience.

Contents:

Some preliminary considerations regarding a musical semantics – An ontogenetic perspective on musical cognition – A view from cognitive semantics – Neuromusicology and the musical mind – Emotion and music – The construction of musical meaning – Some musical elements and their cognitive responses – The embodiment of musical form – Signification in music.

The Author:

Ole Kühl (b. 1950) has spent most of his life working as a musician-cum-composer with improvised music like jazz, fusion and world music. Recently, he turned to the academic field, where he has specialized in a semiotic approach to musicology.



Bern, Berlin, Bruxelles,
Frankfurt am Main, New York,
Oxford, Wien, 2007. 261 pp.

*European Semiotics: Language,
Cognition, and Culture. Vol. 7*
Edited by Per Aage Brandt,
Wolfgang Wildgen, and
Barend van Heusden

ISBN 978-3-03911-282-1 pb.

sFr. 69.- / €* 47.60 / €** 49.- /
€ 44.50 / £ 28.90 / US-\$ 57.95

Please send your order to:

Peter Lang AG

European Academic Publishers
Moosstrasse 1 · P.O. Box 350
CH-2542 Pieterlen · Switzerland

Tel. +41 (0)32 376 17 17
Fax +41 (0)32 376 17 27

e-mail: info@peterlang.com
Internet: www.peterlang.com